Review

Machine Learning Models Using Wearable and Digital Phenotyping Data for Attention-Deficit/Hyperactivity Disorder Assessment: Systematic Review and Diagnostic Test Accuracy Meta-Analysis

Qianhui Wen[1,2], MD; Qian Wang[1,2], MD; Rong Luo[1,2], MD

[1]Department of Pediatrics, West China Second University Hospital of Sichuan University, Chengdu, China

[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, West China Second University Hospital of Sichuan University, Chengdu, Sichuan, China

Corresponding Author:

Qian Wang, MD
Department of Pediatrics
West China Second University Hospital of Sichuan University
No. 20, Section 3, Renmin South Road, Wuhou District
Chengdu 610041
China
Phone: 86 18883903831
Email: 448316346@qq.com

Abstract

Background: Attention-deficit/hyperactivity disorder (ADHD) assessment requires integration of developmental history, impairment, psychosocial context, and information from multiple informants and settings. Wearable and digital phenotyping data may provide objective behavioral, physiological, and sensor-derived signals relevant to ADHD assessment, but the diagnostic performance and clinical readiness of machine learning (ML) and deep learning (DL) models using these data remain uncertain.

Objective: This review evaluated ML/DL models using wearable or digital phenotyping data for ADHD assessment and examined heterogeneity across study, data source, population, algorithm, and validation characteristics.

Methods: We searched PubMed, Embase, Web of Science, the Cochrane Library, and IEEE Xplore from inception through January 23, 2026. Eligible full-text reports evaluated ML/DL classifiers using objective data derived from wearable devices or digital platforms in participants with ADHD and comparators. We excluded nonoriginal reports and studies lacking ML/DL models or eligible objective digital inputs. Reports providing eligible participant-level 2×2 estimates entered quantitative synthesis, with 1 estimate retained per cohort; others were summarized qualitatively. Risk of bias was assessed with QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies–AI). Bayesian bivariate Reitsma models estimated pooled sensitivity and specificity, with a secondary HSROC (hierarchical summary receiver operating characteristic) analysis. Descriptive subgroup analyses examined the validation approach, data sources, population, and algorithms.

Results: After 254 duplicates were removed, 616 records were screened. A total of 34 eligible reports were included, with report-level sample sizes totaling 10,476 participants (3402 with ADHD and 7074 comparators). Of these, 17 reports contributed 1 retained participant-level 2×2 estimate per cohort to the quantitative synthesis, and 17 were summarized qualitatively. Data sources were wearable devices (n=18), digital task–based assessments (n=8), other or multimodal recordings (n=7), and smartphone passive sensing (n=1); 17 reports involved laboratory-based assessments, whereas 17 involved ambulatory or real-world data collection. Pooled sensitivity was 0.875 (95% credible interval [CrI] 0.808-0.927), and pooled specificity was 0.904 (95% CrI 0.840-0.948); the secondary HSROC area under the receiver operating characteristic curve was 0.936 (95% CrI 0.880-0.970). Pooled positive and negative likelihood ratios were 9.153 (95% CrI 5.255-17.041) and 0.138 (95% CrI 0.079-0.219), respectively. The Deeks funnel-plot asymmetry test was considered inconclusive because of limited power (*P*=.99). Certainty of evidence was very low for both pooled sensitivity and pooled specificity. Reference standards varied, and patient-selection risk was high in all 17 quantitative reports and 16 of 17 qualitative reports.

Conclusions: ML/DL models showed high apparent discrimination in selected datasets; however, case-control designs, healthy controls, heterogeneous reference standards, limited external validation, and high patient-selection risk limit transportability to routine clinical practice. They should remain investigational adjuncts, not stand-alone diagnostic tools. Progress toward clinical use will require external validation in clinically representative cohorts, assessment of model calibration, and evidence of incremental clinical value within ADHD assessment pathways.

Trial Registration: PROSPERO CRD420261301888; https://www.crd.york.ac.uk/PROSPERO/view/CRD420261301888

J Med Internet Res 2026;28:e101227; doi: 10.2196/101227

Keywords: attention-deficit/hyperactivity disorder; ADHD; digital phenotyping; wearable electronic devices; machine learning; deep learning; AI; diagnostic accuracy; systematic review; meta-analysis

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a common neurodevelopmental disorder with onset typically in childhood and persists for many individuals into adolescence and adulthood [1,2]. It is associated with substantial impairment across academic, social, emotional, and later occupational functioning [1,2]. ADHD is also clinically heterogeneous, with symptom presentation varying across developmental stages and contexts [1-3].

ADHD assessment remains fundamentally a clinical process, particularly in child and adolescent practice [1,2,4,5]. Current guidelines recommend a comprehensive evaluation that integrates developmental and psychiatric history, psychosocial context, and information from multiple settings and informants [4-6]. Rating scales may serve as useful adjuncts but are insufficient on their own to establish the diagnosis [4-6]. In routine care, diagnostic decision-making is often complicated by discrepancies across informants, context-dependent variation in symptoms and impairment, and multistep, sometimes prolonged, assessment pathways [4-7]. Accordingly, new tools should be judged not only by diagnostic discrimination but also by their ability to support clinical decision-making within standard care pathways.

These challenges have contributed to growing interest in objective measures that may complement, rather than replace, standard clinical assessment [8-10]. Candidate diagnostic biomarkers in neurodevelopmental disorders have often shown limited replication and uncertain clinical readiness [10]. Digital measures should therefore be judged not only by their apparent discrimination but also by their validity, generalizability, and clinical utility. Advances in digital psychiatry and digital phenotyping have made it increasingly feasible to capture activity, sleep, physiological arousal, attention, and behavioral variability through wearables, smartphones, remote monitoring platforms, eye-tracking systems, and computerized task platforms, including repeated or continuous measurements in ecologically relevant settings [8,11-16]. In this review, we use the term "wearable or digital phenotyping data" to refer to objectively captured digital signals that do not rely on self-report and that are used as inputs for machine learning (ML) or deep learning (DL) models. Such measures may be especially relevant to ADHD because core symptoms fluctuate over time and across settings and may be difficult to characterize fully during a single clinic visit [1,2,5,6,8,14].

In parallel, ML and DL methods have become increasingly prominent in ADHD research [17,18]. These approaches are well suited to analyzing high-dimensional and multimodal data, including movement patterns, physiological signals, gaze metrics, neurocognitive task features, and derived digital biomarkers [12,14,17-20]. Early studies have reported encouraging discrimination between ADHD and non-ADHD groups, raising the possibility that these approaches could support more objective and scalable assessment [12,14,19,20]. However, the translational significance of these approaches remains uncertain [17,18].

The literature nevertheless remains difficult to interpret as a whole. Studies vary substantially in sampling strategies, diagnostic reference standards, control selection, signal modalities, feature engineering, algorithm choice, and validation approaches [17,18]. In addition, many prediction models rely primarily on internal rather than external validation and may therefore be vulnerable to overfitting and to limited transportability [17,18,21]. Reported performance may consequently be optimistic relative to real-world clinical settings, and the clinical utility of these approaches remains to be established [17,18,21].

This caution is consistent with evidence from earlier generations of objective ADHD tools [22-26]. Continuous performance tests (CPTs) have shown limited diagnostic utility when used in isolation [22,23], and the Quantified Behavioral Test (QbTest) is better understood as an adjunct within a broader clinical pathway than as a stand-alone screening or diagnostic instrument [24-26]. Wearable and digital objective measures may offer advantages in accessibility, ecological reach, and repeated assessments, particularly in child and adolescent psychiatry [8,11,14].

Accordingly, we conducted a systematic review and meta-analysis to evaluate the diagnostic accuracy of ML/DL models based on wearable or digital phenotyping data for supporting ADHD assessment. Our aims were to estimate overall diagnostic accuracy, including pooled sensitivity, specificity, and summary discrimination; to examine heterogeneity across study and model characteristics; and to assess the extent to which the current evidence supports the clinical use of these models within standard ADHD assessment pathways.

Methods

Registration and Methodological Amendments

This systematic review and diagnostic test accuracy (DTA) meta-analysis was registered in PROSPERO (International Prospective Register of Systematic Reviews) on February 8, 2026 (CRD420261301888), and reported in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) guideline [27]. No separate full protocol was prepared. During subsequent dataset auditing, the one-independent-estimate-per-cohort rule and the estimate-selection hierarchy were formalized as postregistration methodological amendments; they did not alter the review question, target condition, or overall eligibility scope. These postregistration methodological amendments are summarized in Table S1 in Multimedia Appendix 1. The completed PRISMA-DTA checklist is provided as Checklist 1.

Eligibility Criteria

We included original full-text reports of studies enrolling participants of any age with ADHD and a comparator group and evaluating ML/DL classifiers using objective behavioral, physiological, or sensor-derived data captured through wearable or digital platforms. Eligible index tests included passive sensing, task-based digital assessments, and multimodal digitally captured signals used as non–self-report inputs for classification.

Reports providing a participant-level 2×2 diagnostic accuracy estimate, either reported directly or reconstructible from published data, were considered for quantitative DTA synthesis; final estimate selection followed the procedures described under "Data Synthesis and Statistical Analysis." Other eligible reports were summarized qualitatively.

We excluded nonoriginal or non–full-text publications, including reviews, conference abstracts, editorials, and commentaries; reports that did not evaluate ML/DL methods; and reports that did not use objective data derived from wearable devices or digital platforms. Only full-text reports published in English were eligible.

Search Strategy and Information Sources

We systematically searched PubMed, Embase, Web of Science, the Cochrane Library, and IEEE Xplore from database inception through January 23, 2026. The search strategy combined controlled vocabulary terms and free-text terms related to ADHD, wearable and digital data capture, and ML or DL methods. No geographic restrictions were applied. The full search strategies for all databases are provided in Supplement 1 in Multimedia Appendix 1.

Study Selection and Data Extraction

Records were imported into EndNote (Clarivate) and deduplicated. Two reviewers (Q Wen and Q Wang) independently screened titles and abstracts and assessed full-text reports for eligibility. Disagreements were resolved through discussion, with arbitration by a third reviewer (RL) when needed. Reasons for full-text exclusion were recorded and are shown in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram ("Results" section); report-level exclusion reasons are provided in Table S2 in Multimedia Appendix 1.

Potentially overlapping reports were linked at the cohort level; quantitative estimate selection followed the procedures described under "Data Synthesis and Statistical Analysis," whereas eligible companion reports were used for supplementary information and were retained for qualitative synthesis.

Two reviewers independently extracted and cross-checked data using a pilot-tested form. Extracted items included author, year, region, reported sample size, age group, reference-standard procedures, data source and signal characteristics, model and validation characteristics, and diagnostic accuracy and performance data.

Reference-standard information was coded at the report level as (1) the diagnostic framework or named diagnostic instrument, when reported, such as *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders* [Fourth Edition]), *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]), *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*), K-SADS-PL (Kiddie Schedule for Affective Disorders and Schizophrenia–Present and Lifetime Version), or MINI (Mini-International Neuropsychiatric Interview); and (2) the reference-standard type was classified as one of the following: structured or semistructured diagnostic interview, clinician diagnosis with limited procedural detail, chart-based or electronic medical record diagnosis, composite ascertainment, rating-scale–based ascertainment, insufficiently reported ascertainment, or not reported. For reports reanalyzing previously assembled datasets, coding was harmonized to the source-dataset reference standard only when the parent dataset documentation clearly described case ascertainment; otherwise, ascertainment was classified as insufficiently reported.

Data and model characteristics included the data source type, signal modality, induction task, collection setting and duration, algorithm category, and validation strategy, including whether external validation was performed. We also descriptively coded the comparator spectrum for all 34 eligible reports using 4 categories: healthy/neurotypical/normative or nonclinical non-ADHD controls; non-ADHD general, school/community, or cohort comparators; other clinical/differential or mixed clinical/neurodevelopmental comparators; and insufficiently reported. Report-level coding is presented in Table S3 in Multimedia Appendix 1. For reports summarized qualitatively, we also reviewed the reported sampling frame and noted whether it clearly reflected a referral-based ADHD differential-diagnostic pathway.

Risk of Bias and Applicability Assessment

Two reviewers (Q Wen and Q Wang) independently assessed risk of bias and applicability concerns using QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies–AI) across 4 domains: patient selection, index test, reference standard, and flow and timing [28]. Each signaling question was rated as "yes," "no," or "unclear," and disagreements were resolved through discussion, with arbitration by a third reviewer (RL) when needed. QUADAS-AI results were summarized separately for the quantitative and qualitative subsets, and reference-standard heterogeneity was additionally summarized descriptively.

Certainty of Evidence

Certainty of evidence for the primary pooled sensitivity and specificity outcomes was assessed using the GRADE diagnostic test accuracy (GRADE-DTA) framework across domains of risk of bias, indirectness, inconsistency, imprecision, and publication bias. Detailed judgments and downgrade rationales are provided in Table S4 in Multimedia Appendix 1.

Data Synthesis and Statistical Analysis

For quantitative reports, participant-level 2×2 tables were either extracted directly or reconstructed from published sensitivity, specificity, and group sample sizes using standard methods [29]. Study-level sensitivity and specificity were calculated with exact 95% CIs using the Clopper-Pearson method [30]. The Bayesian Reitsma and HSROC (hierarchical summary receiver operating characteristic) models used the raw integer 2×2 counts and handled zero cells through the binomial likelihood without continuity correction. Zero cells occurred in Oh et al [31] (false positive [FP]=0), Amado-Caballero et al [32] (FP=0), and Kam et al [33] (false negative=0). For the Deeks test only, 0.5 was added to all 4 cells of these reports for diagnostic odds ratio (DOR)–derived calculations; the effective sample size (ESS) was calculated from the original uncorrected margins.

When multiple eligible estimates were available within a retained cohort, selection followed the postregistration hierarchy of external validation, separate fixed hold-out testing, nested cross-validation, simple cross-validation, and other resampling procedures. Within the same validation tier, the reported primary model, threshold, or end point was preferred; otherwise, the clinically most relevant estimate with the most complete diagnostic accuracy data was retained.

The primary synthesis used a Bayesian bivariate random-effects Reitsma model with binomial likelihoods, logit links, and a noncentered parameterization [34]. Pooled sensitivity and specificity were reported as posterior medians with 95% credible intervals (CrIs); positive and negative likelihood ratios, DOR, between-study variability, and posterior prediction intervals (PIs) were derived from the posterior distribution. Conventional inverse-variance study weights were not applicable to this model and were therefore not reported. A secondary Bayesian HSROC model was fitted to the same 2×2 data to characterize the sensitivity-specificity trade-off and estimate a summary AUC (area under the receiver operating characteristic curve) [35].

Descriptive subgroup analyses examined validation strategy, data source, population, and algorithm type. Separate fixed hold-out test sets and external validation were classified as independent test-set validation, whereas cross-validation and other resampling procedures were classified as internal validation. Validation-strategy classification was used for descriptive characterization and subgrouping and did not determine eligibility for the primary overall DTA synthesis; reports with unclear validation were displayed but not pooled within the validation-strategy subgroup analysis. Subgroup levels with fewer than 3 studies were displayed but not pooled, and no formal interaction testing or meta-regression was performed because of limited study numbers and sparse strata. Comparator spectrum and reference-standard type were coded descriptively but were not pooled as meta-analytic subgroups. Post hoc pooling was not performed because these strata were sparse and uneven and were strongly associated with other study characteristics, including population, data modality, algorithm type, and validation strategy, making between-stratum differences highly susceptible to ecological confounding. Detailed stratum distributions and the rationale are provided in Supplement 2.7 in Multimedia Appendix 1.

Small-study effects were assessed using the Deeks funnel plot asymmetry test, with study-level ln(DOR) regressed on $1/\sqrt{ESS}$ and a 2-sided threshold of $P<.10$ [36]. Sensitivity analyses comprised leave-one-out refitting, a wider-prior Reitsma analysis, Deeks analyses excluding studies with the highest ln(DOR) values or largest absolute weighted residuals, and a small-sample analysis excluding reports with fewer than 20 participants in either the ADHD or comparator arm, based on study-level enrollment data reported for each study. After these exclusions, the primary Reitsma model was refitted, and the Deeks test was repeated in the remaining 12 studies.

Analyses were conducted in Python. Detailed model specifications, priors, sampling settings, random seeds, software versions, and convergence diagnostics are provided in Supplement 2 in Multimedia Appendix 1.

Ethical Considerations

This review used publicly available data only and did not require ethics approval.

Results

Study Characteristics

After removal of 254 duplicates, 616 records underwent title and abstract screening, 68 reports were assessed for eligibility, and 34 eligible reports were included in the review (Figure 1). Of these, 17 reports contributed one retained participant-level 2×2 estimate per cohort to the quantitative synthesis, whereas 17 were summarized qualitatively because

they did not contribute a separately retained participant-level 2×2 estimate owing to incomplete reporting, overlapping cohorts, or nonindependent analytic units. Characteristics of the eligible reports are summarized in Table 1 [12-14,19,31-33,37-63]. Additional details on study design, participant recruitment or presentation, and funding sources are provided in Table S5 in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of report selection for the systematic review and diagnostic test accuracy (DTA) meta-analysis.

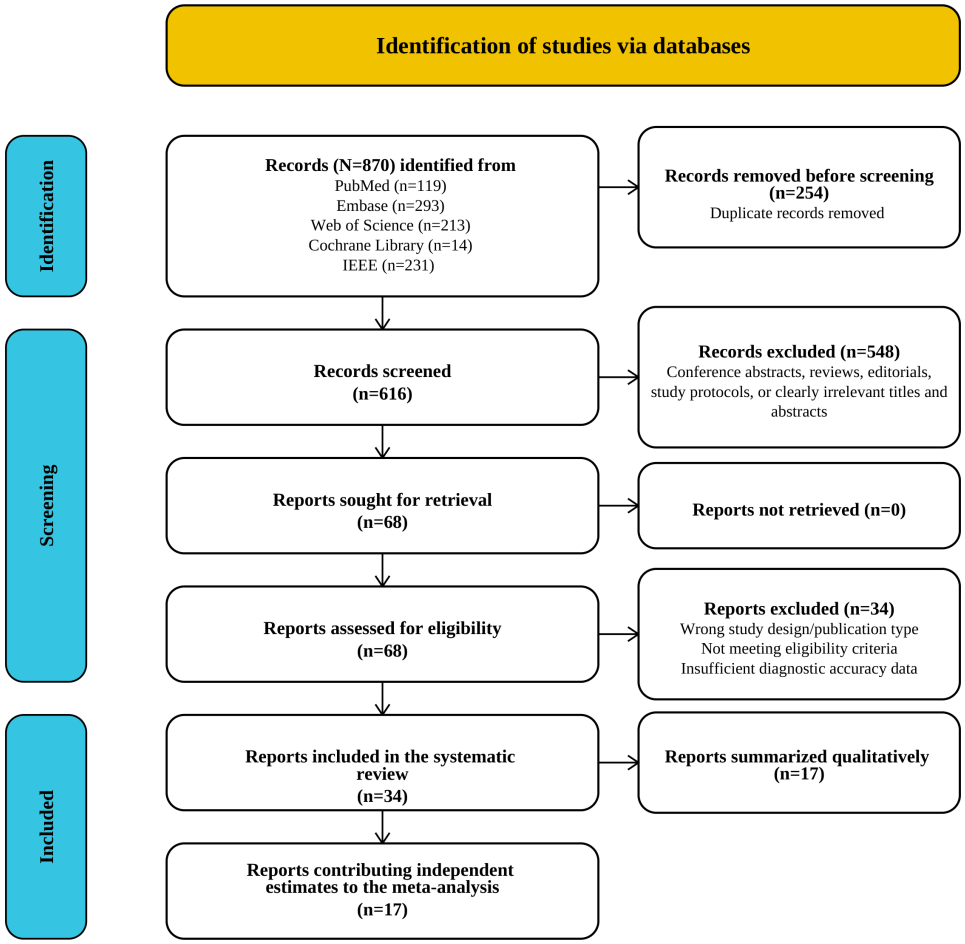


Table 1. Characteristics of included reports assessing wearable and digital phenotyping data for attention-deficit/hyperactivity disorder (ADHD) assessment.

Author (year)	Region	Population	Diagnostic framework/ version	Reference standard type	Sample size, N (ADHD/ comparator)[a]	Data source type	Signal modalities	Induction task	Data setting	Collection duration	Algorithm type	Validation method
Reports included in the quantitative synthesis (n=17)												
Zhang et al [37], 2026	China	Children	Not explicitly reported	Clinician diagnosis with limited procedural detail	94 (44/50)	Digital task (CPT[b]/VR[c])	CPT/ Simon/Go-No-go task performance	Combined continuous screening paradigm (CCSP)	Laboratory	4 min	ML[d]	Hold-out
Casals et al [38], 2025	Sweden	Children and adults	Not explicitly reported	Clinician diagnosis with limited procedural detail	1244 (292/952)	Digital task (CPT/VR)	CPT performance; face tracking; smartphone IMU[e]	CPT	Laboratory	10 min	ML	5-fold CV[f]+hold-out
Castillo-Gutierrez [39], 2025	Norway	Adults	*DSM-IV*[g], *ICD-10*[h]	Structured or semistructured diagnostic interview	77 (45/32)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	7 h optimal/ multiple days	ML	LOSO-CV[i]
Jiang et al [14], 2024	China	Children	Not explicitly reported	Chart-based or electronic medical record diagnosis	30 (17/13)	Wearable	Accelerometer/activity; heart rate; sleep patterns	Passive monitoring	Ambulatory/real-world	12 wk	ML	5-fold CV
Oh et al [31], 2024	South Korea	Children	Not explicitly reported	Composite ascertainment	20 (15/5)	Digital task (CPT/VR)	VR movement tracking; controller interaction logs; eye tracking; voice audio	VR daily-life scenarios/tasks	Laboratory	20-25 min	DL[j]	Unclear
Andrikopoulos et al [12], 2024	Greece	Adults	*DSM-5*[k]	Composite ascertainment	76 (32/44)	Other/ multimodal	EDA[l]; HR[m]/HRV[n]; skin temperature	Stroop test	Laboratory	15 min	ML	100-fold stratified CV
Wiebe et al [19], 2024	Germany	Adults	*DSM-5*	Structured or semistructured diagnostic interview	86 (43/43)	Digital task (CPT/VR)	VR-CPT[o] performance; head actigraphy; eye tracking; EEG[p]	VR-CPT	Laboratory	54 min	ML	5-fold nested CV + external validation
Loftness et al [40], 2023	United States	Children	K-SADS-PL[q]/ preschool version	Composite ascertainment	104 (20/84)	Digital task (CPT/VR)	Accelerometer; gyroscope; audio	Mood-induction tasks	Laboratory	7 min	ML	Stratified 2-fold CV

Author (year)	Region	Population	Diagnostic framework/ version	Reference standard type	Sample size, N (ADHD/ comparator)[a]	Data source type	Signal modalities	Induction task	Data setting	Collection duration	Algorithm type	Validation method
Lee et al [41], 2023	South Korea	Children	Not explicitly reported	Composite ascertainment	828 (66/762)	Other/ multimodal	RGB-D[r] skeletal tracking	Robot-led ADHD screening game	Laboratory	1 h	DL	LOSO-CV + hold-out
Zhang et al [42], 2023	China	Children	*DSM-5*, K-SADS-PL	Structured or semistructured diagnostic interview	40 (20/20)	Wearable	Fingertip force	VR-based fingertip force control task	Laboratory	19-28 min	ML	10-fold CV
Lindhiem et al [13], 2022	United States	Children	K-SADS-PL	Composite ascertainment	30 (15/15)	Wearable	Smartwatch motion; GPS; heart rate; Bluetooth	Passive monitoring	Ambulatory/real-world	2 d	ML	LOSO-CV
Slobodin et al [43], 2020	Israel	Children	*DSM*[s]-based clinical diagnosis	Clinician diagnosis with limited procedural detail	458 (213/245)	Digital task (CPT/VR)	CPT performance	MOXO-CPT[t] task	Laboratory	15.2 min	ML	100-fold CV
Amado-Caballero et al [32], 2020	Spain	Children	*DSM-5*	Clinician diagnosis with limited procedural detail	148 (73/75)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	24 h	DL	10-fold CV
O'Mahony et al [44], 2014	Spain	Children	*DSM-IV-TR*[u]	Clinician diagnosis with limited procedural detail	43 (24/19)	Wearable	IMU accelerometer; gyroscope	Waiting, consultant, and CPT	Laboratory	1 h	ML	LOOCV[v]
Teicher et al [45], 2012	United States	Adults	*DSM-IV*	Composite ascertainment	100 (40/60)	Other/ multimodal	Infrared motion capture	Cognitive control task, cued response task, embedded memory task, Conners CPT-II[w]	Laboratory	20 min	ML	Repeated random-split CV
Martín-Martínez et al [46], 2012	Spain	Children	*DSM-IV*	Clinician diagnosis with limited procedural detail	63 (31/32)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	24 h	ML	LOOCV
Kam et al [33], 2010	South Korea	Children	*DSM-IV*	Composite ascertainment	142 (10/132)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	3 h	ML	10-fold CV
Reports included in the qualitative synthesis (n=17)												
Ji et al [47], 2025	United States	Adults	Not explicitly reported	Insufficiently reported ascertainment	103 (51/52)	Wearable	Activity and HRV	Passive monitoring	Ambulatory/real-world	6.6 (SD 1.3) d (ADHD), 7.2 (SD 0.9) d (controls)	ML	10-fold CV
Ware et al [48], 2025	United States	Adults	*DSM-5*	Composite ascertainment	12 (8/4)	Smartphone	Smartphone app usage logs	Passive monitoring	Ambulatory/real-world	10 wk	ML	5-fold CV

Author (year)	Region	Population	Diagnostic framework/ version	Reference standard type	Sample size, N (ADHD/ comparator)[a]	Data source type	Signal modalities	Induction task	Data setting	Collection duration	Algorithm type	Validation method
Rahman [49], 2025	United States	Children	K-SADS-5[x]/ *DSM-5*	Structured or semistructured diagnostic interview	450 (225/225)	Wearable	Sedentary time; resting heart rate; energy expenditure	Passive monitoring	Ambulatory/real-world	21 d	ML	10-fold CV + hold-out
Doctor et al [50], 2025	United States	Children and adults	Not explicitly reported	Clinician diagnosis with limited procedural detail	75 (41/34)	Wearable	Triaxial linear acceleration; tri-axial angular velocity; orientation angles	Reaching paradigm	Laboratory	15 min	DL	5-fold CV + hold-out
Liu et al [51], 2025	United States	Children	K-SADS-5/ *DSM-5*	Structured or semistructured diagnostic interview	1953 (216/1737)	Wearable	Heart rate; calories; activity intensity; steps; METs[y]; sleep level; sleep intensity	Passive monitoring	Ambulatory/real-world	NR[z]	DL	Hold-out
Feretzakis et al [52], 2025	Greece	Adults	Not explicitly reported	Insufficiently reported ascertainment	45 (23/22)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	NR	ML	5-fold CV + hold-out
von Polier et al [53], 2025	Germany	Adults	*DSM-IV-TR*	Composite ascertainment	591 (387/204)	Other/ multimodal	Speech audio	Free speech and counting	Laboratory	3 min	ML	10-fold CV
Wang et al [54], 2025	China	Children	Not explicitly reported	Rating-scale–based ascertainment	50 (13/37)	Other/ multimodal	Wrist accelerometer; clinical tabular variables	Passive monitoring	Ambulatory/real-world	NR	DL	10-fold CV + hold-out
Akash et al [55], 2025	India	Adults	Source-dataset dependent/not redescribed	Insufficiently reported ascertainment	116 (45/71)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	NR	DL	Hold-out
Li et al [56], 2024	United Kingdom	Adults	*DSM-5*	Clinician diagnosis with limited procedural detail	22 (10/12)	Other/ multimodal	Speech audio; text transcripts	DIVA[aa] interview	Laboratory	8 min	ML	LOSO-CV
Li et al [57], 2024	United Kingdom	Adults	Not explicitly reported	Clinician diagnosis with limited procedural detail	22 (10/12)	Other/ multimodal	RGB[ab] video; pose/ skeleton keypoints	DIVA interview, CANTAB[ac], beep reaction task, and watching videos	Laboratory	1-1.5 h	DL	Hold-out

Author (year)	Region	Population	Diagnostic framework/ version	Reference standard type	Sample size, N (ADHD/ comparator)[a]	Data source type	Signal modalities	Induction task	Data setting	Collection duration	Algorithm type	Validation method
Shafna et al [58], 2024	India	Adults	Not explicitly reported	Insufficiently reported ascertainment	116 (45/71)	Wearable	Wrist actigraphy/ motor activity	Passive monitoring	Ambulatory/real-world	2-3 d	ML	10-fold CV + hold-out
Kim et al [59], 2023	United States	Children	K-SADS[ad]/ *DSM-5*	Structured or semistructured diagnostic interview	1090 (79/1011)	Wearable	Wrist actigraphy; heart rate; sleep	Passive monitoring	Ambulatory/real-world	21 d	ML	Hold-out
Amado-Caballero et al [60], 2023	Spain	Children	ADHD-RS-IV[ae]/*DSM-5*	Composite ascertainment	139 (70/69)	Wearable	Wrist actigraphy/ tri-axial acceleration magnitude	Passive monitoring	Ambulatory/real-world	approximately 24 h	DL	10-fold CV
Kaur & Kahlon [61], 2022	India	Adults	Not explicitly reported	Insufficiently reported ascertainment	103 (51/52)	Wearable	Wrist actigraphy	Passive monitoring	Ambulatory/real-world	24 h	ML	10-fold CV
Yeh et al [62], 2020	China	Children	Not explicitly reported	Composite ascertainment	68 (37/31)	Digital task (CPT/VR)	VR task performance; HMD[af] head rotation	CPT and audio test	Laboratory	20 min	ML	5-fold CV

Author (year)	Region	Population	Diagnostic framework/ version	Reference standard type	Sample size, N (ADHD/ comparator)[a]	Data source type	Signal modalities	Induction task	Data setting	Collection duration	Algorithm type	Validation method
Lee et al [63], 2008	South Korea	Children	*DSM-IV*	Composite ascertainment	1938 (1091/847)	Digital task (CPT/VR)	CPT task performance	CPT	Laboratory	NR	ML	Hold-out

[a]N indicates the total sample size; values in parentheses indicate the numbers of participants with ADHD and comparators, respectively.
[b]CPT: continuous performance test.
[c]VR: virtual reality.
[d]ML: machine learning.
[e]IMU: inertial measurement unit.
[f]CV: cross-validation.
[g]*DSM-IV*: *DSM* (Fourth Edition).
[h]*ICD-10*: *International Statistical Classification of Diseases, Tenth Revision*.
[i]LOSO-CV: leave-one-subject-out cross-validation.
[j]DL: deep learning.
[k]*DSM-5*: *DSM* (Fifth Edition).
[l]EDA: electrodermal activity.
[m]HR: heart rate.
[n]HRV: heart rate variability.
[o]VR-CPT: virtual reality CPT.
[p]EEG: electroencephalography.
[q]K-SADS-PL: Kiddie Schedule for Affective Disorders and Schizophrenia–Present and Lifetime Version.
[r]RGB-D: red-green-blue plus depth.
[s]*DSM*: *Diagnostic and Statistical Manual of Mental Disorders*.
[t]MOXO-CPT: MOXO Continuous Performance Test.
[u]*DSM-IV-TR*: *DSM-IV* (Text Revision).
[v]LOOCV: leave-one-out cross-validation.
[w]CPT-II: Conners Continuous Performance Test II.
[x]K-SADS-5: K-SADS for DSM-5.
[y]METs: metabolic equivalents.
[z]NR: not reported.
[aa]DIVA: Diagnostic Interview for ADHD in Adults.
[ab]RGB: red-green-blue.
[ac]CANTAB: Cambridge Neuropsychological Test Automated Battery.
[ad]K-SADS: Kiddie Schedule for Affective Disorders and Schizophrenia.
[ae]ADHD-RS-IV: ADHD Rating Scale, Fourth Edition.
[af]HMD: head-mounted display.

Across the 34 reports published between 2008 and 2026, report-level sample sizes totaled 10,476 participants, including 3402 with ADHD and 7074 comparators. Because some reports drew on overlapping source cohorts, this total does not represent a deduplicated count of unique individuals. The age distribution showed that 19 reports focused on children, 13 on adults or adult-source cohorts, and 2 included both age groups. The most frequently represented countries were the United States (n=9), China (n=5), Spain (n=4), South Korea (n=4), and India (n=3). Data sources included wearable devices (n=18), digital task–based assessments, such as CPT and virtual reality (VR) paradigms (n=8), other or multimodal physiological or behavioral recordings (n=7), and smartphone-derived passive sensing data (n=1). Data collection settings were evenly divided between laboratory and ambulatory or real-world environments (n=17 reports each). Algorithms were classified as ML in 25 reports and DL in 9. Among the 17 quantitative reports, 4 used independent test-set validation, 12 used cross-validation or other internal resampling procedures, and 1 had an unclear validation strategy.

Across all 34 reports, 18 used healthy/neurotypical/normative or nonclinical non-ADHD controls; 5 used non-ADHD general, school/community, or cohort comparators; 10 used other clinical/differential or mixed clinical/neurodevelopmental comparators; and 1 was insufficiently reported. Among the 17 quantitative reports, the corresponding counts were 12, 1, 4, and 0, respectively. Thus, only 4 of the 17 quantitative reports involved clinical/differential or mixed clinical/neurodevelopmental comparators. Only 1 quantitative report clearly described referral-based sampling, and no included report clearly documented a consecutive, prospectively assembled referral cohort representative of routine ADHD differential diagnosis.

Reference Standard Heterogeneity

Among the 34 eligible reports, reference-standard types were composite ascertainment (n=12), clinician diagnosis with limited procedural detail (n=9), structured or semistructured diagnostic interview (n=6), insufficiently reported ascertainment (n=5), chart-based or electronic medical record diagnosis (n=1), and rating-scale–based ascertainment (n=1). Among the 17 quantitative reports, the corresponding distribution was composite ascertainment (n=7), clinician diagnosis with limited procedural detail (n=6), structured or semistructured diagnostic interview (n=3), and chart-based or electronic medical record diagnosis (n=1). Thus, the pooled estimates summarize performance across heterogeneous ADHD ascertainment procedures rather than against a single, uniform clinical reference standard.

Calibration Reporting

Only 1 of the 34 (2.9%) included reports provided a formal calibration-related assessment: Zhang et al [37] reported a Hosmer-Lemeshow goodness-of-fit test for their logistic screening model (χ_8^2=0.697, P>.99). No included report presented a calibration plot, calibration intercept or slope, Brier score, or expected calibration error. Calibration performance could not, therefore, be meaningfully synthesized across the evidence base.

Qualitative Synthesis

Across the 17 reports summarized qualitatively, the reported diagnostic performance generally showed moderate-to-good apparent discrimination [47-63], but pooling was not appropriate because of heterogeneity in data modalities, task context, sample composition, validation strategy, comparator spectrum, analytic unit, and reporting completeness. Most comparator groups were not clearly representative of routine referral-based ADHD differential diagnosis.

Reports using wearable or smartphone-derived data examined motor activity, rest-activity patterns, sleep or circadian features, app use, and related behavioral or physiological signals across ambulatory, classroom, and laboratory settings [47-52,54,55,58-61]. The remaining reports examined voice- or interview-derived acoustic and text features, video- or skeleton-based action recognition, VR-based or other structured behavioral tasks, and computerized CPT-based novelty detection [53,56,57,62,63]. Further details, including comparator-spectrum coding and reported performance metrics, are provided in Supplement 3 and Tables S3 and S6 in Multimedia Appendix 1.

Quantitative Synthesis

Pooled Sensitivity, Specificity, and HSROC Summary

Using the same 17 retained quantitative estimates, the primary Reitsma model yielded a pooled sensitivity of 0.875 (95% CrI 0.808-0.927) and a specificity of 0.904 (95% CrI 0.840-0.948; Figure 2), whereas the secondary HSROC model yielded a summary AUC of 0.936 (95% CrI 0.880-0.970; Figure 3).

Figure 2. Study-level and pooled estimates of (A) sensitivity and (B) specificity [12-14,19,31-33,37-46]. CrI: credible interval.

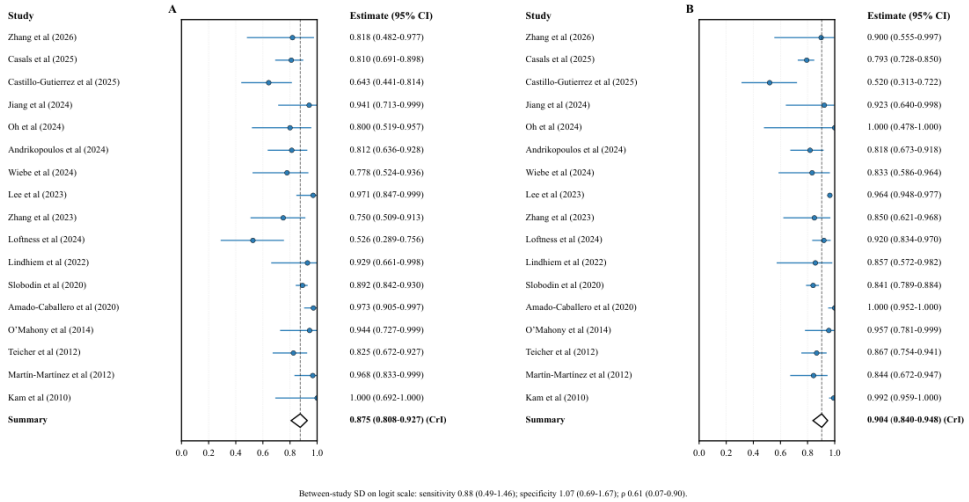
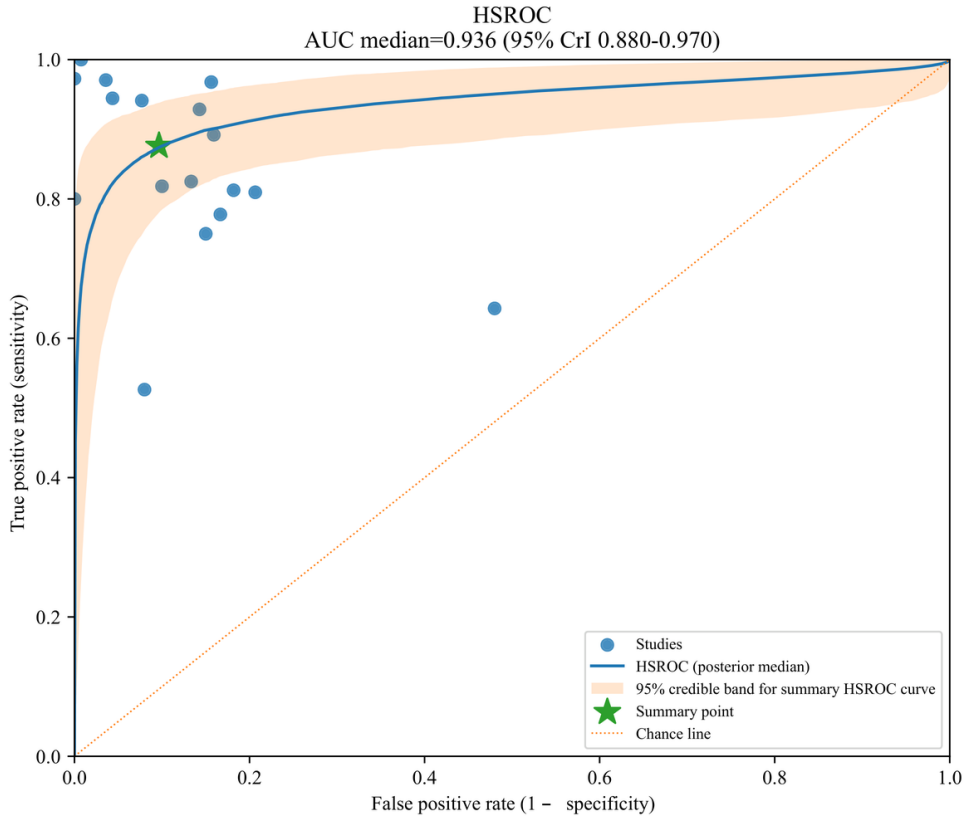


Figure 3. Secondary HSROC (hierarchical summary receiver operating characteristic) curve summarizing diagnostic performance across studies. AUC: area under the receiver operating characteristic curve.



Between-study heterogeneity in the bivariate model was reflected by logit-scale SDs of 0.882 (95% CrI 0.491-1.461) for sensitivity and 1.067 (95% CrI 0.690-1.673) for specificity, with an estimated correlation of 0.613 (95% CrI 0.072-0.905) between the corresponding study-level random effects. Posterior-derived secondary accuracy measures were a positive likelihood ratio of 9.153 (95% CrI 5.255-17.041), a negative likelihood ratio of 0.138 (95% CrI 0.079-0.219), and a DOR of 67.446 (95% CrI 26.237-182.043; Table 2). For a new comparable study, the posterior predictive median sensitivity was 0.876 (95% PI 0.508-0.982) and the specificity was 0.904 (95% PI 0.483-0.991). The wide PIs indicate substantial expected variation in diagnostic performance across comparable future studies and show that performance in a new setting could differ considerably from the pooled estimates.

Table 2. Posterior-derived secondary accuracy measures, between-study variability, and posterior predictive study-level performance from the Reitsma model.[a]

Measure or parameter	Values
A. Posterior-derived likelihood ratios and DOR[b], posterior median (95% CrI[c])	
Measure: PLR[d]	9.153 (5.255-17.041)
Measure: NLR[e]	0.138 (0.079-0.219)
Measure: DOR	67.446 (26.237-182.043)
B. Between-study variability (random effects on the logit scale), posterior median (95% CrI)	
Parameter: SD (logit sensitivity)	0.882 (0.491-1.461)
Parameter: SD (logit specificity)	1.067 (0.690-1.673)
ϱ[f]	0.613 (0.072-0.905)
C. Posterior predictive performance for a new comparable study, posterior median (95% PI[g])	
Measure: predicted sensitivity	0.876 (0.508-0.982)
Measure: predicted specificity	0.904 (0.483-0.991)

[a]Panel A reports posterior-derived PLR, NLR, and DOR as medians with 95% CrIs. Panel B reports logit-scale between-study SDs for sensitivity and specificity and their correlation. Panel C reports posterior predictive study-level sensitivity and specificity as medians with 95% PIs; these intervals reflect both posterior uncertainty and between-study heterogeneity.
[b]DOR: diagnostic odds ratio.
[c]CrI: credible interval.
[d]PLR: positive likelihood ratio.
[e]NLR: negative likelihood ratio.
[f]Correlation coefficient.
[g]PI: prediction interval.

Subgroup Analyses

Validation Strategy

Independent test-set validation (separate hold-out testing or external validation; k=4) yielded a pooled sensitivity of 0.856 (95% CrI 0.701-0.946) and a specificity of 0.892 (95% CrI 0.697-0.963), whereas cross-validation or other internal resampling procedures (k=12) yielded a pooled sensitivity of 0.884 (95% CrI 0.794-0.944) and a specificity of 0.907 (95% CrI 0.820-0.957). Oh et al [31] had an unclear validation strategy and was displayed but not pooled (Figure S1 in Multimedia Appendix 1).

Data Source

Digital task–based studies, including CPT- or VR-based paradigms (k=6), yielded a pooled sensitivity of 0.795 (95% CrI 0.629-0.896) and a specificity of 0.855 (95% CrI 0.782-0.928); wearable-based studies (k=8) yielded a sensitivity of 0.918 (95% CrI 0.818-0.968) and a specificity of 0.920 (95% CrI 0.792-0.975); and other or multimodal studies (k=3) yielded a sensitivity of 0.873 (95% CrI 0.656-0.961) and a specificity of 0.901 (95% CrI 0.685-0.967; Figure S2 in Multimedia Appendix 1).

Population

Studies in children (k=12) yielded a pooled sensitivity of 0.908 (95% CrI 0.825-0.960) and a specificity of 0.939 (95% CrI 0.884-0.972), whereas studies in adults (k=4) yielded a pooled sensitivity of 0.768 (95% CrI 0.620-0.868) and a specificity of 0.778 (95% CrI 0.533-0.901). One study included both children and adults and was displayed but not pooled (Figure S3 in Multimedia Appendix 1).

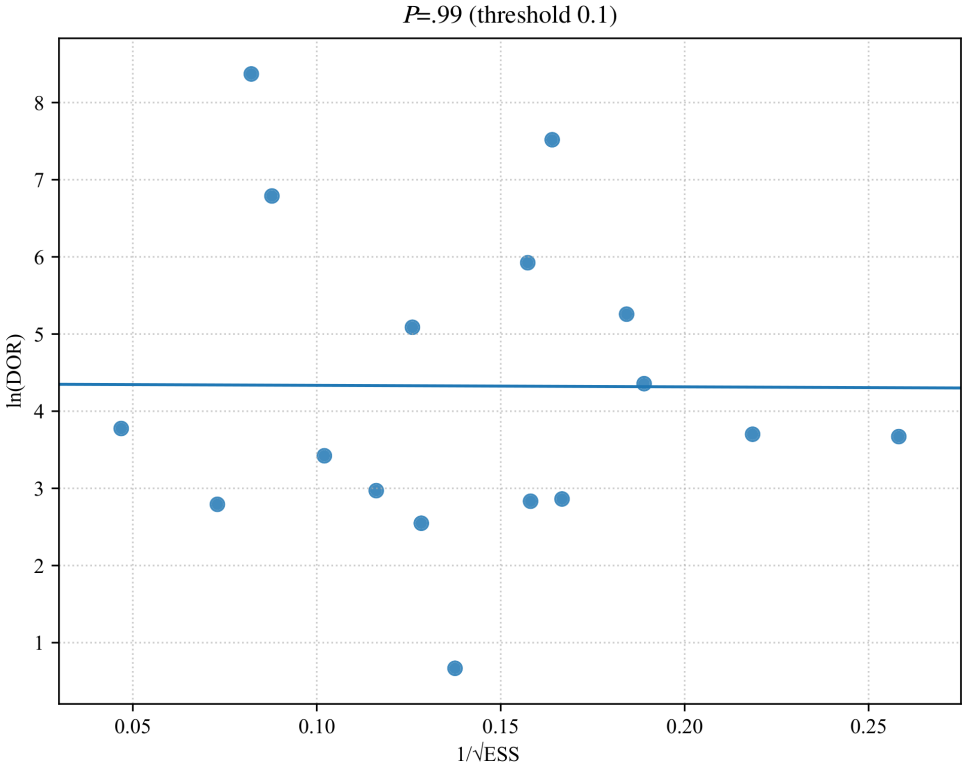
Algorithm Type

Traditional ML models (k=14) yielded a pooled sensitivity of 0.849 (95% CrI 0.779-0.914) and a specificity of 0.874 (95% CrI 0.803-0.928), whereas DL models (k=3) yielded a pooled sensitivity of 0.941 (95% CrI 0.737-0.983) and a specificity of 0.972 (95% CrI 0.883-0.995; Figure S4 in Multimedia Appendix 1).

Small-Study Effects and Sensitivity Analyses

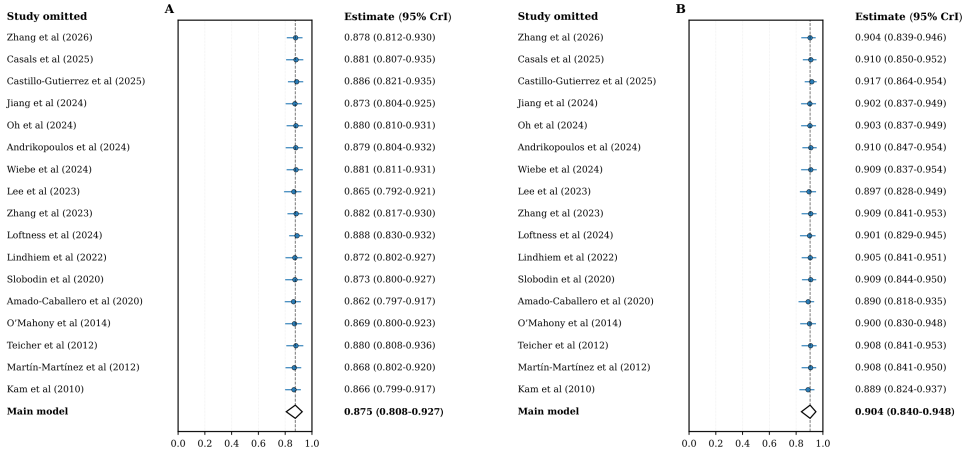
The Deeks funnel-plot asymmetry test yielded a slope of –0.195 (95% CI –22.848 to 22.457; *P*=.99; Figure 4). Given only 17 quantitative estimates and the wide CI, the result was considered inconclusive; the nonsignificant *P* value does not exclude publication bias or selective reporting. The lowest DOR (DOR=1.95) point was Castillo-Gutierrez et al [39]. Sensitivity analyses excluding studies with the highest ln(DOR) values or largest absolute weighted residuals were likewise nonsignificant (*P*=.59 to .95; Table S7 and Figure S5 in Multimedia Appendix 1) and remained underpowered.

Figure 4. The Deeks funnel plot asymmetry test for small-study effects. DOR: diagnostic odds ratio; ESS: effective sample size.



In the small-sample sensitivity analysis excluding reports with fewer than 20 participants in either the ADHD or comparator arm (k=12 retained), pooled sensitivity was 0.851 (95% CrI 0.755-0.916) and specificity was 0.876 (95% CrI 0.785-0.932), decreases of 0.024 and 0.028, respectively, from the primary estimates. The Deeks funnel-plot asymmetry test in this subset yielded a slope of –7.820 (95% CI –43.050 to 27.411; *P*=.63) and remained inconclusive because of limited power. Across the 17 leave-one-out model refits, the posterior median pooled sensitivity estimates ranged from 0.862 to 0.888, and the corresponding specificity estimates ranged from 0.889 to 0.917. The individual estimates and their 95% CrIs are presented in Figure 5. Although no single report materially altered the pooled summary estimates, leave-one-out stability does not address systematic limitations shared across the evidence base, including patient-selection bias, comparator-spectrum bias, and heterogeneity in reference standards.

Figure 5. Leave-one-out influence analyses for (A) pooled sensitivity and (B) pooled specificity [12-14,19,31-33,37-46]. Each row shows the pooled estimate after omission of the named study; the dashed line and diamond indicate the main model. CrI: credible interval.

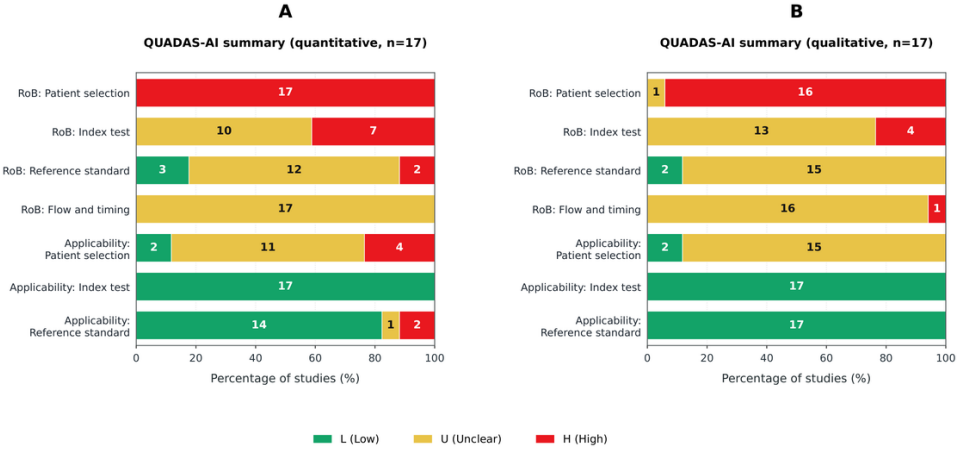


Risk of Bias and Applicability

QUADAS-AI results were summarized separately for the quantitative (n=17; Figure 6A) and qualitative (n=17; Figure 6B) subsets. Patient-selection risk was high in all 17 quantitative reports and 16 of 17 qualitative reports. Across both subsets, index-test risk was never low, flow-and-timing risk was unclear or high in every report, and reference-standard risk was unclear or high in 14 quantitative and 15 qualitative reports. Applicability concerns were lower overall: index-test applicability was of low concern in all reports, reference-standard applicability was of low concern in 14 quantitative reports and all 17 qualitative reports, whereas

patient-selection applicability was unclear or high in 15 reports in each subset. Study-level QUADAS-AI risk-of-bias and applicability judgments for all 34 included reports are provided in Table S8 in Multimedia Appendix 1.

Figure 6. QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies–AI) summary of risk of bias (RoB) and applicability assessments for reports contributing to the quantitative synthesis (n=17) and reports summarized qualitatively (n=17). Numbers within the bars indicate the numbers of reports.



Certainty of Evidence

Using GRADE-DTA, certainty of evidence for both pooled sensitivity and pooled specificity was rated very low, with downgrades for very serious risk of bias, serious indirectness, and serious inconsistency. No separate imprecision downgrade was applied because clinically meaningful sensitivity or specificity thresholds were not prespecified and a post hoc threshold could double-count heterogeneity. Publication bias was not formally downgraded because the Deeks analyses were inconclusive rather than demonstrably asymmetric, although residual concern remained. Detailed judgments and downgrade rationales are provided in Table S4 in Multimedia Appendix 1.

Discussion

In this systematic review and meta-analysis, ML/DL models based on wearable or digital phenotyping data showed high apparent discrimination for ADHD in selected research datasets. The primary Reitsma model yielded a pooled sensitivity of 0.875 (95% CrI 0.808-0.927) and a specificity of 0.904 (95% CrI 0.840-0.948), whereas the secondary HSROC model yielded a summary AUC of 0.936 (95% CrI 0.880-0.970). These findings do not establish readiness for stand-alone diagnosis. Many studies used case-control samples, healthy or normative comparators, and internal validation, whereas routine ADHD assessment requires distinguishing ADHD from overlapping neurodevelopmental, emotional, behavioral, and sleep-related presentations [17,18,21]. This mismatch leaves the evidence base vulnerable to spectrum bias and optimistic performance estimates. Consistent with these limitations, GRADE-DTA certainty for both pooled sensitivity and specificity was very low. Until clinically representative external validation demonstrates generalizability and incremental clinical value, these models should remain investigational adjuncts to comprehensive ADHD assessment [4,5,22-26].

All subgroup findings should be regarded as exploratory and hypothesis-generating. Study characteristics were strongly interrelated, and sparse strata, together with the absence of formal interaction testing, preclude attributing observed differences to any single factor or interpreting them as genuine causal or performance differences.

Wearable-based studies had higher pooled point estimates than digital task–based studies. This descriptive pattern may partly reflect differences in measurement context: wearable sensing can capture motor activity, sleep-related patterns, and behavioral variability over longer or more naturalistic periods [11,64], whereas CPT- or VR-based paradigms primarily sample performance during structured testing sessions [22,23]. However, the data modality was closely associated with age group, sample size, algorithm type, validation strategy, and comparator spectrum, so the observed subgroup differences cannot be attributed to modality alone.

Age-group findings should also be interpreted cautiously. Pediatric studies reported higher pooled point estimates than adult studies, but this descriptive pattern may reflect developmental differences in ADHD presentation as well as differences in sampling and study design [3,65]. Hyperactivity, impulsivity, and attentional dysregulation may be more externally observable in children [3,66], whereas adult ADHD is often more heterogeneous and complicated by comorbidity [65,67-69]. Given the smaller adult subgroup and between-study variability, the observed difference should not be interpreted as an independent age-group effect.

Limited external validation raises concerns about data leakage and optimistic performance estimates. Leakage can occur when preprocessing, feature engineering or selection, threshold selection, or hyperparameter tuning are informed by data later used for evaluation; repeated model development and selection within the same dataset can further inflate performance [70]. TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI), PROBAST+AI (Prediction

Model Risk of Bias Assessment Tool+AI), and STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies–AI) [71-73] emphasize clear separation of model development and evaluation, transparent reporting of preprocessing and feature selection, appropriate validation, and prespecified or independently evaluated thresholds. These requirements highlight gaps in the present evidence base: only 4 of 17 quantitative reports used an independent test set, and only 1 explicitly reported external validation, while reporting was often insufficient to determine whether leakage risks had been adequately controlled. Accordingly, the similar pooled estimates for internal-validation and independent-test-set subgroups should not be interpreted as evidence that internal validation provides equivalent transportable clinical performance.

Although the DL subgroup had higher point estimates than the traditional ML subgroup, it comprised only 3 studies with wide CrIs, and no formal between-subgroup interaction test was performed. Because algorithm type could not be disentangled from other study characteristics, the higher estimates should be viewed as descriptive rather than as evidence of an algorithm-type effect.

The pooled diagnostic metrics aggregate estimates generated from heterogeneous ADHD ascertainment procedures rather than from a single, uniform clinical reference standard. Accordingly, pooled sensitivity and specificity should be interpreted as summaries across heterogeneous diagnostic criteria, not as performance against a common clinical gold standard.

The comparator spectrum also raises concerns about spectrum bias and limits clinical transportability. Among the 17 quantitative reports, 12 used healthy or other non-clinical comparators, compared with 4 using clinical/differential or mixed clinical/neurodevelopmental comparators and 1 using a general school/community comparator. Comparisons with clearly separated healthy controls can inflate apparent discrimination relative to routine clinical practice, where ADHD must be distinguished from competing neurodevelopmental, emotional, behavioral, sleep-related, and other psychiatric presentations [17,18,21]. The pooled estimates may therefore overstate performance in real-world referral and differential-diagnostic settings.

These concerns are compounded by substantial between-study heterogeneity and wide posterior PIs, indicating that performance in a future comparable study could differ considerably from the pooled estimates. QUADAS-AI assessments further showed pervasive patient-selection risk and frequent reference-standard concerns across both quantitative and qualitative evidence. Both the primary Deeks analysis and the analysis in the k=12 small-sample subset were inconclusive because of limited power and wide CIs; their nonsignificant results do not exclude small-study effects, publication bias, or selective outcome reporting. Excluding the 5 reports meeting the small-sample criterion modestly reduced pooled sensitivity and specificity without materially changing the overall interpretation or addressing systematic design limitations. Taken together, these findings indicate that the pooled estimates may be optimistic and should not be assumed to generalize to routine care.

Despite these limitations, the review incorporated several methodological safeguards, including PROSPERO registration, cohort-level linkage of potentially overlapping reports, retention of 1 estimate per cohort, and prioritization of independent test-set estimates when available. The Bayesian bivariate framework accounted for between-study variability and was complemented by HSROC, posterior predictive, prior-sensitivity, leave-one-out, and small-sample analyses. These safeguards strengthen the robustness of the synthesis but cannot overcome systematic limitations of the underlying evidence base. Overall, the field remains at an early translational stage rather than being ready for routine implementation [74,75].

In practice, ML/DL models using wearable or digital phenotyping data may complement comprehensive ADHD assessment. Potential applications include cases in which symptom reports are incomplete, limited to a single setting, or difficult to reconcile across informants, particularly when repeated or naturalistic measurements may add information about day-to-day functioning [6,64,76]. Other potential roles include triage, longitudinal monitoring, and support when standard assessment leaves residual diagnostic uncertainty [25,26,77,78]. However, the pooled likelihood ratios do not establish sufficient diagnostic certainty for stand-alone use, particularly in referral settings with substantial comorbidity and diagnostic overlap [5,6,79,80]. The key translational question is whether these models improve triage efficiency or provide clinically useful information beyond existing ADHD assessment procedures [26,77].

Clinical usefulness also depends on the model's intended role and decision threshold. A model used to prioritize referrals, prompt further assessment, or monitor change involves different false-positive and false-negative trade-offs from one intended to support diagnostic confirmation. Discrimination should therefore be considered alongside calibration, prespecified thresholds, and decision-analytic measures such as net benefit. A high AUC alone does not establish good calibration or clinical usefulness across sites, devices, populations, or referral pathways [71,72,81,82].

Calibration assessment was nearly absent, with only 1 of 34 reports providing any formal calibration-related assessment. Future diagnostic AI studies should report calibration alongside discrimination, particularly during external validation, using calibration plots and quantitative measures such as calibration-in-the-large, calibration slope, and the Brier score, where appropriate [81].

Future studies should address clearly defined clinical questions in prospectively assembled, clinically representative cohorts that include common differential diagnoses and comorbidities rather than relying on small case-control or healthy-control comparisons [17,21,78,80]. External validation across sites, devices, and populations should be routine [17,21,71,72]. Investigators should prespecify the intended clinical role and threshold strategy, report threshold-specific sensitivity, specificity, and predictive values at

clinically plausible prevalence levels, and evaluate net benefit where appropriate [6,71,72,82]. Reporting should transparently document preprocessing, feature selection, hyperparameter tuning, leakage-prevention procedures, missing-data handling, and subgroup performance [70-73]. Ultimately, models should be evaluated within ADHD assessment pathways for incremental clinical value, including effects on clinical decisions, waiting times, downstream assessment burden, and patient-centered outcomes, rather than on discrimination alone [26,74,75,77].

In summary, ML/DL models based on wearable or digital phenotyping data showed high apparent discrimination for ADHD in selected research samples, but the evidence remains preliminary and does not establish performance in routine clinical practice. These models should remain investigational adjuncts rather than stand-alone diagnostic tools. Progress toward clinical use will require external validation in representative cohorts, assessment of calibration, and demonstration of incremental clinical value within real-world ADHD assessment pathways.

Generative AI Use

During the preparation and revision of this manuscript, the authors used the generative AI (GenAI) tools ChatGLM (GLM-5; Zhipu AI) and ChatGPT (GPT-5.4; OpenAI) to assist with language refinement, revision planning, and manuscript-formatting suggestions. These GenAI tools were not used to generate original study data, conduct primary study screening or data extraction, independently determine statistical results, or verify references. The authors reviewed and edited all AI-assisted content and take full responsibility for the final content of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The report-level information required to interpret the systematic review is provided in the article and Multimedia Appendix 1. The quantitative data were extracted or reconstructed from previously published reports cited in the reference list. No new participant-level data were generated or collected for this review.

Authors' Contributions

Conceptualization: Q Wen, Q Wang, RL
Data curation: Q Wen, Q Wang
Formal analysis: Q Wen
Investigation: Q Wen, Q Wang
Methodology: Q Wen, Q Wang, RL
Supervision: RL
Visualization: Q Wen
Writing – original draft: Q Wen
Writing – review & editing: Q Wen, Q Wang, RL
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategies; statistical methods, model specifications, diagnostics, and sensitivity analyses; qualitative synthesis; postregistration methodological amendments; comparator-spectrum coding; GRADE-DTA (GRADE approach for diagnostic test accuracy) certainty assessment; study design and funding characteristics; excluded full-text articles; and supplementary tables and figures.
[DOCX File (Microsoft Word File), 1793 KB-Multimedia Appendix 1]

Checklist 1

PRISMA-DTA checklist.
[PDF File (Adobe File), 102 KB-Checklist 1]

References

1. Faraone SV, Bellgrove MA, Brikell I, et al. Attention-deficit/hyperactivity disorder. Nat Rev Dis Primers. Feb 22, 2024;10(1):11. [doi: 10.1038/s41572-024-00495-0] [Medline: 38388701]
2. Faraone SV, Banaschewski T, Coghill D, et al. The World Federation of ADHD International Consensus Statement: 208 evidence-based conclusions about the disorder. Neurosci Biobehav Rev. Sep 2021;128:789-818. [doi: 10.1016/j.neubiorev.2021.01.022] [Medline: 33549739]
3. Nigg JT, Sibley MH, Thapar A, Karalunas SL. Development of ADHD: etiology, heterogeneity, and early life course. Annu Rev Dev Psychol. Dec 2020;2(1):559-583. [doi: 10.1146/annurev-devpsych-060320-093413] [Medline: 34368774]

4. Wolraich ML, Hagan JF Jr, Allan C, et al. Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics. Oct 2019;144(4):e20192528. [doi: 10.1542/peds.2019-2528] [Medline: 31570648]
5. Attention deficit hyperactivity disorder: diagnosis and management. National Institute for Health and Care Excellence (NICE); 2019. URL: https://www.ncbi.nlm.nih.gov/books/NBK493361/pdf/Bookshelf_NBK493361.pdf [Accessed 2026-11-09]
6. Peterson BS, Trampush J, Brown M, et al. Tools for the diagnosis of ADHD in children and adolescents: a systematic review. Pediatrics. Apr 1, 2024;153(4):e2024065854. [doi: 10.1542/peds.2024-065854] [Medline: 38523599]
7. Bonati M, Cartabia M, Zanetti M, Lombardy ADHD Group. Waiting times for diagnosis of attention-deficit hyperactivity disorder in children and adolescents referred to Italian ADHD centers must be reduced. BMC Health Serv Res. Sep 18, 2019;19(1):673. [doi: 10.1186/s12913-019-4524-0] [Medline: 31533711]
8. Knyazhansky M, Shrot T. ADHD diagnostic tools across ages: traditional and digital approaches. Front Psychiatry. 2025;16:1668070. [doi: 10.3389/fpsyt.2025.1668070] [Medline: 41164091]
9. Tan YL, Ma G, McIntyre RS, et al. The clinical utility of the objective measures for diagnosing and monitoring attention-deficit and hyperactivity disorder (ADHD) in adults: a systematic review. Psychiatry Int. 2026;7(1):1. [doi: 10.3390/psychiatryint7010001]
10. Cortese S, Solmi M, Michelini G, et al. Candidate diagnostic biomarkers for neurodevelopmental disorders in children and adolescents: a systematic review. World Psychiatry. Feb 2023;22(1):129-149. [doi: 10.1002/wps.21037] [Medline: 36640395]
11. Sankesara H, Denyer H, Sun S, et al. Identifying digital markers of attention-deficit/hyperactivity disorder (ADHD) in a remote monitoring setting: prospective observational study. JMIR Form Res. Jan 29, 2025;9:e54531. [doi: 10.2196/54531] [Medline: 39885373]
12. Andrikopoulos D, Vassiliou G, Fatouros P, Tsirmpas C, Pehlivanidis A, Papageorgiou C. Machine learning-enabled detection of attention-deficit/hyperactivity disorder with multimodal physiological data: a case-control study. BMC Psychiatry. Aug 5, 2024;24(1):547. [doi: 10.1186/s12888-024-05987-7] [Medline: 39103819]
13. Lindhiem O, Goel M, Shaaban S, et al. Objective measurement of hyperactivity using mobile sensing and machine learning: pilot study. JMIR Form Res. Apr 25, 2022;6(4):e35803. [doi: 10.2196/35803] [Medline: 35468089]
14. Jiang Z, Chan AYL, Lum D, et al. Wearable signals for diagnosing attention-deficit/hyperactivity disorder in adolescents: a feasibility study. JAACAP Open. 2024;3(4):875-889. [doi: 10.1016/j.jaacop.2024.11.003] [Medline: 41367997]
15. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. Jun 2016;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
16. Yoo JH, Kang C, Lim JS, et al. Development of an innovative approach using portable eye tracking to assist ADHD screening: a machine learning study. Front Psychiatry. 2024;15:1337595. [doi: 10.3389/fpsyt.2024.1337595] [Medline: 38426003]
17. Salazar de Pablo G, Iniesta R, Bellato A, et al. Individualized prediction models in ADHD: a systematic review and meta-regression. Mol Psychiatry. Dec 2024;29(12):3865-3873. [doi: 10.1038/s41380-024-02606-5] [Medline: 38783054]
18. Zhao C, Xu Y, Li R, Li H, Zhang M. Artificial intelligence in ADHD assessment: a comprehensive review of research progress from early screening to precise differential diagnosis. Front Artif Intell. 2025;8:1624485. [doi: 10.3389/frai.2025.1624485] [Medline: 40964142]
19. Wiebe A, Selaskowski B, Paskin M, et al. Virtual reality-assisted prediction of adult ADHD based on eye tracking, EEG, actigraphy and behavioral indices: a machine learning analysis of independent training and test samples. Transl Psychiatry. Dec 31, 2024;14(1):508. [doi: 10.1038/s41398-024-03217-y] [Medline: 39741130]
20. Liu Z, Li J, Zhang Y, et al. Auxiliary diagnosis of children with attention-deficit/hyperactivity disorder using eye-tracking and digital biomarkers: case-control study. JMIR mHealth uHealth. Nov 29, 2024;12:e58927. [doi: 10.2196/58927] [Medline: 39477792]
21. Meehan AJ, Lewis SJ, Fazel S, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol Psychiatry. Jun 2022;27(6):2700-2708. [doi: 10.1038/s41380-022-01528-4] [Medline: 35365801]
22. Arrondo G, Mulraney M, Iturmendi-Sabater I, et al. Systematic review and meta-analysis: clinical utility of continuous performance tests for the identification of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. Feb 2024;63(2):154-171. [doi: 10.1016/j.jaac.2023.03.011] [Medline: 37004919]
23. Varela JL, Magnante AT, Miskey HM, Ord AS, Eldridge A, Shura RD. A systematic review of the utility of continuous performance tests among adults with ADHD. Clin Neuropsychol. Oct 2024;38(7):1524-1585. [doi: 10.1080/13854046.2024.2315740] [Medline: 38424025]

24. Bellato A, Hall CL, Groom MJ, et al. Practitioner review: clinical utility of the QbTest for the assessment and diagnosis of attention-deficit/hyperactivity disorder - a systematic review and meta-analysis. J Child Psychol Psychiatry. Jun 2024;65(6):845-861. [doi: 10.1111/jcpp.13901] [Medline: 37800347]
25. Tomlinson E, Owen-Smith A, Benavente M, et al. QbTest for ADHD assessment and medication management: a mixed-methods systematic review of impact on clinical outcomes and patient, carer and clinician experiences. BMJ Open. Apr 17, 2025;15(4):e095479. [doi: 10.1136/bmjopen-2024-095479] [Medline: 40246562]
26. Digital technologies for assessing attention deficit hyperactivity disorder (ADHD). National Institute for Health and Care Excellence (NICE); Oct 21, 2024. URL: https://www.nice.org.uk/guidance/htg729/resources/digital-technologies-for-assessing-attention-deficit-hyperactivity-disorder-adhd-pdf-1809600650784709 [Accessed 2026-03-17]
27. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. Aug 14, 2020;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
28. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. Oct 2021;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
29. Bossuyt PM, Deeks JJ, Leeflang MM, Takwoingi Y, Flemyng E. Evaluating medical tests: introducing the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Cochrane Database Syst Rev. Jul 20, 2023;7(7):ED000163. [doi: 10.1002/14651858.ED000163] [Medline: 37470764]
30. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-413. [doi: 10.1093/biomet/26.4.404]
31. Oh S, Joung YS, Chung TM, et al. Diagnosis of ADHD using virtual reality and artificial intelligence: an exploratory study of clinical applications. Front Psychiatry. 2024;15:1383547. [doi: 10.3389/fpsyt.2024.1383547] [Medline: 38887727]
32. Amado-Caballero P, Casaseca-de-la-Higuera P, Alberola-Lopez S, et al. Objective ADHD diagnosis using convolutional neural networks over daily-life activity records. IEEE J Biomed Health Inform. Sep 2020;24(9):2690-2700. [doi: 10.1109/JBHI.2020.2964072] [Medline: 31905156]
33. Kam HJ, Shin YM, Cho SM, Kim SY, Kim KW, Park RW. Development of a decision support model for screening attention-deficit hyperactivity disorder with actigraph-based measurements of classroom activity. Appl Clin Inform. 2010;1(4):377-393. [doi: 10.4338/ACI-2010-05-RA-0033] [Medline: 23616848]
34. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. Oct 2005;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
35. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. Oct 15, 2001;20(19):2865-2884. [doi: 10.1002/sim.942] [Medline: 11568945]
36. 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] [Medline: 16085191]
37. Zhang Y, Cao D, Hou H, et al. Enhancing ADHD early screening with a mobile serious game: a combined continuous screening paradigm. IEEE Trans Games. 2026;18(2):275-284. [doi: 10.1109/TG.2026.3650962]
38. Casals N, Larsson S, Hansen M. Machine learning on a smartphone-based CPT for ADHD prediction. Front Psychiatry. 2025;16:1564351. [doi: 10.3389/fpsyt.2025.1564351] [Medline: 41280456]
39. Castillo-Gutierrez PA. Managing data uncertainty and machine learning for adult ADHD classification using accelerometry: OBF-psychiatric case study. medRxiv. Preprint posted online on Aug 27, 2025. [doi: 10.1101/2025.08.26.25332257]
40. Loftness BC, Halvorson-Phelan J, OLeary A, et al. The ChAMP App: a scalable mHealth technology for detecting digital phenotypes of early childhood mental health. IEEE J Biomed Health Inform. Nov 29, 2024;28(4):2304-2313. [doi: 10.1109/JBHI.2023.3337649]
41. Lee DW, Lee SH, Ahn DH, Lee GH, Jun K, Kim MS. Development of a multiple RGB-D sensor system for ADHD screening and improvement of classification performance using feature selection method. Appl Sci. 2023;13(5):2798. [doi: 10.3390/app13052798]
42. Zhang Z, Kang S, Yu J, et al. Quantitative identification of ADHD tendency in children with immersive fingertip force control tasks. IEEE Trans Neural Syst Rehabil Eng. 2023;31:4561-4569. [doi: 10.1109/TNSRE.2023.3332467] [Medline: 37955997]
43. Slobodin O, Yahav I, Berger I. A machine-based prediction model of ADHD using CPT data. Front Hum Neurosci. 2020;14:560021. [doi: 10.3389/fnhum.2020.560021] [Medline: 33093829]

44. O'Mahony N, Florentino-Liano B, Carballo JJ, Baca-García E, Rodríguez AA. Objective diagnosis of ADHD using IMUs. Med Eng Phys. Jul 2014;36(7):922-926. [doi: 10.1016/j.medengphy.2014.02.023] [Medline: 24657100]
45. Teicher MH, Polcari A, Fourligas N, Vitaliano G, Navalta CP. Hyperactivity persists in male and female adults with ADHD and remains a highly discriminative feature of the disorder: a case-control study. BMC Psychiatry. Nov 7, 2012;12(1):190. [doi: 10.1186/1471-244X-12-190] [Medline: 23134619]
46. Martín-Martínez D, Casaseca-de-la-Higuera P, Alberola-López S, et al. Nonlinear analysis of actigraphic signals for the assessment of the attention-deficit/hyperactivity disorder (ADHD). Med Eng Phys. Nov 2012;34(9):1317-1329. [doi: 10.1016/j.medengphy.2011.12.023] [Medline: 22297088]
47. Ji Y, Zhang-Lea J, Tran J. Automated ADHD detection using dual-modal sensory data and machine learning. Med Eng Phys. May 2025;139:104328. [doi: 10.1016/j.medengphy.2025.104328] [Medline: 40306880]
48. Ware S, Rijal K, Knouse LE. SmartADHDMonitor: a novel approach to automatic ADHD monitoring through smartphone app usage data. IEEE Annu Comput Softw Appl Conf (COMPSAC). 2025:854-865. [doi: 10.1109/COMPSAC65507.2025.00113]
49. Rahman MM. Unlocking the potential of wearable technology: Fitbit-derived measures for predicting ADHD in adolescents. Front Child Adolesc Psychiatry. 2025;4:1504323. [doi: 10.3389/frcha.2025.1504323] [Medline: 40487008]
50. Doctor KP, McKeever C, Wu D, et al. Deep learning diagnosis plus kinematic severity assessments of neurodivergent disorders. Sci Rep. Jul 8, 2025;15(1):20269. [doi: 10.1038/s41598-025-04294-9] [Medline: 40628787]
51. Liu JJ, Borsari B, Li Y, et al. Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. Cell. Jan 23, 2025;188(2):515-529. [doi: 10.1016/j.cell.2024.11.012] [Medline: 39706190]
52. Feretzakis G, Boufeas I, Fotakidis S, et al. Enhanced machine learning approach to ADHD classification using actigraphy data. Stud Health Technol Inform. Jun 26, 2025;328:136-140. [doi: 10.3233/SHTI250688] [Medline: 40588896]
53. von Polier GG, Ahlers E, Volkening J, et al. Exploring voice as a digital phenotype in adults with ADHD. Sci Rep. May 24, 2025;15(1):18076. [doi: 10.1038/s41598-025-01989-x] [Medline: 40413201]
54. Wang L, Yang G. ViT-BiLSTM multimodal learning for paediatric ADHD recognition: integrating wearable sensor data with clinical profiles. Sensors (Basel). Oct 18, 2025;25(20):6459. [doi: 10.3390/s25206459] [Medline: 41157512]
55. AkashSP, ShafnaV, Madhu Kumar SD. Intelligent ADHD detection: leveraging knowledge-enhanced CNNs with low-cost activity sensor data. IEEE Reg 10 Symp (TENSYMP). 2025:1-6. [doi: 10.1109/TENSYMP63728.2025.11144941]
56. Li S, Nair R, Naqvi SM. Acoustic and text features analysis for adult ADHD screening: a data-driven approach utilizing DIVA interview. IEEE J Transl Eng Health Med. 2024;12:359-370. [doi: 10.1109/JTEHM.2024.3369764] [Medline: 38606391]
57. Li Y, Nair R, Naqvi SM. ADHD detection based on human action recognition. Neurosci Appl. 2024;3:104093. [doi: 10.1016/j.nsa.2024.104093] [Medline: 40656101]
58. Shafna V, Madhu Kumar SD. Optimal interval and feature selection in activity data for detecting attention deficit hyperactivity disorder. Comput Biol Med. Sep 2024;179:108909. [doi: 10.1016/j.compbiomed.2024.108909] [Medline: 39053333]
59. Kim WP, Kim HJ, Pack SP, Lim JH, Cho CH, Lee HJ. Machine learning-based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children. JAMA Netw Open. Mar 1, 2023;6(3):e233502. [doi: 10.1001/jamanetworkopen.2023.3502] [Medline: 36930149]
60. Amado-Caballero P, Casaseca-de-la-Higuera P, Alberola-López S, Andrés-de-Llano JM, López-Villalobos JA, Alberola-López C. Insight into ADHD diagnosis with deep learning on actimetry: quantitative interpretation of occlusion maps in age and gender subgroups. Artif Intell Med. Sep 2023;143:102630. [doi: 10.1016/j.artmed.2023.102630] [Medline: 37673587]
61. Kaur A, Kahlon KS. Accurate identification of ADHD among adults using real-time activity data. Brain Sci. Jun 26, 2022;12(7):831. [doi: 10.3390/brainsci12070831] [Medline: 35884638]
62. Yeh SC, Lin SY, Wu EHK, et al. A virtual-reality system integrated with neuro-behavior sensing for attention-deficit/hyperactivity disorder intelligent assessment. IEEE Trans Neural Syst Rehabil Eng. Sep 2020;28(9):1899-1907. [doi: 10.1109/TNSRE.2020.3004545] [Medline: 32746303]
63. Lee HJ, Cho S, Shin MS. Supporting diagnosis of attention-deficit hyperactive disorder with novelty detection. Artif Intell Med. Mar 2008;42(3):199-212. [doi: 10.1016/j.artmed.2007.11.001] [Medline: 18187311]
64. Olinic MS, Stretea R, Chereches C. Wearables in ADHD: monitoring and intervention-where are we now? Diagnostics (Basel). Sep 17, 2025;15(18):2359. [doi: 10.3390/diagnostics15182359] [Medline: 41008731]

65. Cortese S, Bellgrove MA, Brikell I, et al. Attention-deficit/hyperactivity disorder (ADHD) in adults: evidence base, uncertainties and controversies. World Psychiatry. Oct 2025;24(3):347-371. [doi: 10.1002/wps.21374] [Medline: 40948064]
66. Musullulu H. Evaluating attention deficit and hyperactivity disorder (ADHD): a review of current methods and issues. Front Psychol. 2025;16:1466088. [doi: 10.3389/fpsyg.2025.1466088] [Medline: 40066184]
67. Fu X, Wu W, Wu Y, et al. Adult ADHD and comorbid anxiety and depressive disorders: a review of etiology and treatment. Front Psychiatry. 2025;16:1597559. [doi: 10.3389/fpsyt.2025.1597559] [Medline: 40547117]
68. van der Ham M, Bijlenga D, Böhmer M, Beekman ATF, Kooij S. Sleep problems in adults with ADHD: prevalences and their relationship with psychiatric comorbidity. J Atten Disord. Nov 2024;28(13):1642-1652. [doi: 10.1177/10870547241284477] [Medline: 39354860]
69. Waldren LH, Leung FYN, Hargitai LD, et al. Unpacking the overlap between autism and ADHD in adults: a multi-method approach. Cortex. Apr 2024;173:120-137. [doi: 10.1016/j.cortex.2023.12.016] [Medline: 38387375]
70. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun. Feb 28, 2024;15(1):1829. [doi: 10.1038/s41467-024-46150-w] [Medline: 38418819]
71. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
72. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
73. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. Oct 2025;31(10):3283-3289. [doi: 10.1038/s41591-025-03953-8] [Medline: 40954311]
74. Berkhout WEM, van Wijngaarden JJ, Workum JD, et al. Operationalization of artificial intelligence applications in the intensive care unit: a systematic review. JAMA Netw Open. Jul 1, 2025;8(7):e2522866. [doi: 10.1001/jamanetworkopen.2025.22866] [Medline: 40699572]
75. Wang F, Beecy A. Implementing AI models in clinical workflows: a roadmap. BMJ Evid Based Med. Sep 22, 2025;30(5):285-287. [doi: 10.1136/bmjebm-2023-112727] [Medline: 38914450]
76. Bergwerff CE, Buisman RSM, Nibbering N, Noordermeer SDS. Using wearables in mental health care for children and adolescents: a scoping review. Res Child Adolesc Psychopathol. Jan 26, 2026;54(1):16. [doi: 10.1007/s10802-025-01408-9] [Medline: 41588150]
77. Tomlinson E, Ward M, Walker J, et al. Clinical and cost-effectiveness of technologies for the assessment of attention deficit hyperactivity disorder: a systematic review and economic model. Health Technol Assess. Nov 2025;29(58):1-216. [doi: 10.3310/DRDR7171] [Medline: 41220181]
78. Adamou M, Arif M, Asherson P, et al. The adult ADHD assessment quality assurance standard. Front Psychiatry. 2024;15:1380410. [doi: 10.3389/fpsyt.2024.1380410] [Medline: 39156609]
79. Gustafsson U, Larsson S, Casals N, Nolen R, Sanyal RY, Hansen M. Objective assessment of attention deficit hyperactivity disorder with QbMobile: a smartphone application for clinical use. Clin Pract Epidemiol Ment Health. 2025;21(1):e17450179444324. [doi: 10.2174/0117450179444324251118134538] [Medline: 41694709]
80. Young S, Absoud M, Al-Attar Z, et al. The ADHD Assessment Quality Assurance Standard for Children and Teenagers (CAAQAS). Neuropsychiatr Dis Treat. 2024;20:2603-2628. [doi: 10.2147/NDT.S472923] [Medline: 39737064]
81. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. Dec 16, 2019;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
82. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AUC: area under the receiver operating characteristic curve
CPT: continuous performance test
CrI: credible interval
DL: deep learning
DOR: diagnostic odds ratio
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition)

DTA: diagnostic test accuracy
ESS: effective sample size
FP: false positive
GRADE-DTA: GRADE approach for diagnostic test accuracy
HSROC: hierarchical summary receiver operating characteristic
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
K-SADS-PL: Kiddie Schedule for Affective Disorders and Schizophrenia–Present and Lifetime Version
MINI: Mini-International Neuropsychiatric Interview
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies
PROSPERO: International Prospective Register of Systematic Reviews
QbTest: Quantified Behavioral Test
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies–AI
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI
VR: virtual reality

Edited by Ivan Steenstra; peer-reviewed by Alessio Bellato, Wei Zhang; submitted 16.May.2026; initial comments to author 23.Jun.2026; final revised version received 04.Oct.2026; accepted 19.Oct.2026; published 26.Nov.2026

Please cite as:
Wen Q, Wang Q, Luo R
Machine Learning Models Using Wearable and Digital Phenotyping Data for Attention-Deficit/Hyperactivity Disorder Assessment: Systematic Review and Diagnostic Test Accuracy Meta-Analysis
J Med Internet Res 2026;28:e101227
URL: https://www.jmir.org/2026/1/e101227
doi: 10.2196/101227

© Qianhui Wen, Qian Wang, Rong Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Nov.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.