Review

Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis

Jiawen Liu[1,2], MA; Junhui Wang[3], MEng; Zhaobin Wu[3], MEng; Mohamad Ibrani Shahrimin Bin Adam Assim[2], PhD

[1]Youth League Committee, Liuzhou Railway Vocational Technical College, Liuzhou, Guangxi Zhuang Autonomous Region, China
[2]Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Bintulu, Sarawak, Malaysia
[3]School of Automation, Guangxi University of Science and Technology, Liuzhou, Guangxi, China

Corresponding Author:

Jiawen Liu, MA
Youth League Committee
Liuzhou Railway Vocational Technical College
No. 2 Wenyuan Road, Yufeng District
Liuzhou, Guangxi Zhuang Autonomous Region 545000
China
Phone: 60 1116670058
Email: liujiawen@ltzy.edu.cn

Abstract

Background: Artificial intelligence (AI) models have been increasingly explored for predicting treatment response to cognitive behavioral therapy (CBT) in patients with anxiety disorders. Identifying potential responders in advance may help inform treatment planning and support clinical decision-making. Although a growing number of studies have applied AI techniques in this context, reported performance estimates vary across studies, and the overall predictive accuracy has not been comprehensively quantified.

Objective: This systematic review and meta-analysis aims to quantify the overall performance of AI models in predicting treatment response following CBT for anxiety disorders and to examine how data sources, algorithmic approaches, and diagnostic subtypes influence predictive performance.

Methods: A systematic literature search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO up to August 2025. We included studies that validated AI models for predicting CBT treatment response (remission or response) in patients diagnosed with an anxiety disorder. The risk of bias was assessed using the PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) tool. Predictive performance metrics, including sensitivity, specificity, accuracy, and area under the curve (AUC), were extracted and pooled. Pooled estimates for sensitivity, specificity, and diagnostic accuracy were derived using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman method. Prediction intervals were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings.

Results: Eleven studies were included in the meta-analysis. The pooled sensitivity of AI-based models for predicting treatment response was 0.73 (95% CI 0.58-0.85; I^2=82.8%), and the pooled specificity was 0.75 (95% CI 0.59-0.89; I^2=96.7%). The overall pooled accuracy was 0.74 (95% CI 0.62-0.84; I^2=94.6%). The summary AUC was 0.81 (95% CI 0.78-0.85), indicating moderate discriminative performance. Subgroup analyses showed that models incorporating multimodal data achieved superior predictive performance, with a pooled sensitivity of 0.84 and an accuracy of 0.82. In addition, predictive performance was the highest in patients with social anxiety disorder compared with other anxiety disorder subtypes.

Conclusions: This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying social anxiety disorder as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Trial Registration: PROSPERO CRD420251137096; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251137096

J Med Internet Res 2026;28:e86079; doi: 10.2196/86079

Keywords: artificial intelligence; treatment response; anxiety disorders; cognitive behavioral therapy; meta-analysis

Introduction

Anxiety disorders constitute a group of mental disorders characterized by intense, excessive, and persistent worry and fear, exhibiting high prevalence rates. They not only cause significant functional impairment in patients' social, academic, and occupational functioning but also impose substantial socioeconomic burdens at the societal level, including considerable health care resource consumption and productivity losses [1,2]. Cognitive behavioral therapy (CBT) is a first-line evidence-based intervention for anxiety disorders [3,4]. However, treatment response varies considerably, with a notable subset of patients showing suboptimal or inadequate improvement [5]. This heterogeneity underscores a critical clinical need: the ability to preemptively identify likely responders and, more importantly, those at high risk of nonresponse who may benefit from treatment augmentation or alternative interventions from the outset [6]. Early and accurate prediction of treatment response can optimize care pathways, improve resource allocation, and reduce delays, aligning with the goals of precision mental health [7].

Clinical assessment of anxiety disorders typically relies on structured or semistructured interviews and standardized scales [8], such as the Hamilton Anxiety Rating Scale (HAM-A) [9], Clinical Global Impressions–Improvement (CGI-I) [10], and Liebowitz Social Anxiety Scale (LSAS) [7], which are fundamental for symptom identification, outcome evaluation, and longitudinal monitoring. However, they possess inherent limitations for predictive purposes. First, they depend heavily on self-report and subjective ratings, making them vulnerable to recall bias and social desirability effects, thereby compromising objectivity [11]. Second, heterogeneity in measured constructs, scoring procedures, and cutoff thresholds across instruments constrains comparability and diagnostic consistency across studies and clinical settings [12]. Consequently, there is growing recognition that these traditional metrics alone may be insufficient for reliably forecasting an individual's unique response trajectory to a specific therapy such as CBT [13].

Against this background, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has attracted increasing attention as a methodological framework for advancing outcome prediction in mental health research. AI-based models can integrate large-scale, high-dimensional, and multimodal data and capture complex, nonlinear relationships that are difficult to model with conventional statistical techniques [14]. In recent years, a growing number of studies have applied ML and DL methods to predict treatment response across a range of psychiatric interventions, including CBT for anxiety disorders [15,16]. These studies have incorporated diverse data sources, including baseline clinical characteristics, symptom trajectories, neuroimaging markers, psychophysiological measures, and digital behavioral indicators, and have reported varying levels of predictive accuracy [17,18]. However, the anxiety disorders comprise clinically distinct subtypes, such as social anxiety disorder (SAD), generalized anxiety disorder (GAD), and panic disorder (PD), which differ in symptom profiles and underlying mechanisms [19]. This diagnostic heterogeneity may further influence model generalizability and predictive stability across populations [7,20]. To date, however, quantitative evidence synthesizing the predictive performance of AI models for CBT response in different anxiety disorder subtypes remains limited, and potential sources of heterogeneity across studies remain insufficiently explored and require systematic investigation [21].

This systematic review and meta-analysis aims to quantitatively synthesize the overall performance of AI models in predicting treatment response following CBT for anxiety disorders. We also seek to explore the influence of key factors, such as data sources, algorithmic approaches, and diagnostic subtypes, on predictive performance and to preliminarily investigate potential sources of the observed heterogeneity.

Methods

Ethical Considerations

As this is a systematic review and meta-analysis, ethics approval and consent to participate are not applicable. The manuscript does not include the identification images or other personal or clinical details of participants.

The meta-analysis adhered rigorously to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy) reporting guidelines [22] and the PRISMA-S extension for reporting literature searches [23]. We have provided the completed PRISMA-DTA (Checklist 1), PRISMA 2020 Abstract (Checklist 2), and PRISMA-S (Checklist 3) checklists. The research protocol was registered in the PROSPERO registry (registration ID: CRD420251137096).

Feasibility Assessment for Meta-Analysis

Prior to initiating the full systematic review, a preliminary scoping search was conducted to assess the feasibility of quantitative synthesis. This step aimed to determine whether sufficient studies reported the extractable predictive performance metrics necessary for a meta-analysis. In the scoping search, although the field is nascent, several published studies met the core data requirements, supporting the rationale for proceeding with a formal systematic review and meta-analysis.

Search Strategy

A systematic search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO from inception to August 16, 2025. Two independent reviewers (JL and JW) screened titles/abstracts and subsequently assessed full texts, with disagreements resolved by consensus. Search strategies combined controlled vocabulary (MeSH) and free-text terms across 3 concept blocks: AI and methods (eg, “artificial intelligence,” “machine learning,” and “deep learning”), anxiety disorders (eg, “anxiety disorders,” “social anxiety disorder,” “generalized anxiety disorder,” and “panic disorder”), and CBT (eg, “cognitive behavioral therapy,” “CBT,” and “cognitive therapy”). No language or publication year limits were applied at the initial search stage to minimize selection bias. Detailed database-specific strategies are provided in Table S1 in Multimedia Appendix 1. Study selection followed a dual-stage independent screening process. Two reviewers (JL and JW) first screened titles/abstracts and then independently assessed full texts of potentially eligible records. Disagreements were resolved by consensus or by third-party arbitration (ZW). The reference lists of included studies were manually screened for additional records. It should be noted that in conducting this systematic review, we did not contact authors, experts, or manufacturers to obtain additional studies or data, and did not adapt or reuse search strategies from previous reviews. Additionally, a formal peer review of the search strategy was not conducted.

Inclusion and Exclusion Criteria

The PITROS framework was used for the inclusion criteria:

- Participants (P): patients with anxiety disorders receiving CBT, including separation anxiety disorder, selective mutism, specific phobia, SAD, PD/panic attacks, agoraphobia, GAD, and substance/medication-induced anxiety disorder
- Index test (I): AI methods (eg, ML and DL) developed to predict CBT treatment response
- Target condition (T): positive cases defined as clinically significant post-CBT improvement (eg, meaningful symptom reduction); negative cases defined as poor or no response
- Reference standard (R): validated diagnostic/severity scales for anxiety disorders
- Outcomes (O): sensitivity, specificity, area under the curve (AUC), and accuracy
- Setting (S): mental health outpatient clinics, psychiatric hospitals, and clinical trial centers; multicenter prospective or retrospective designs using electronic health records or public databases

The exclusion criteria were as follows: (1) publication in a language other than English at the full-text stage, (2) nonempirical article type (eg, review, case report, conference abstract, meta-analysis, and letter), (3) no assessment of the prediction of CBT response using AI, (4) inclusion of a population that did not have anxiety disorder or that did not undergo CBT, (5) absence of a recognized reference standard or inadequate data to derive the target metrics, and (6) duplicate publication identified using EndNote and manual checking. The database search was unrestricted by language to maximize sensitivity, but non-English publications were excluded during full-text review owing to translation resource constraints. We acknowledge this as a potential limitation, as it may have introduced language bias by excluding relevant studies published in other languages. The independent selection of studies was conducted by 2 reviewers (JL and JW), with disagreements resolved through discussion or by a third reviewer (ZW).

Data Extraction

Two independent reviewers (JL and JW) conducted comprehensive data extraction from full-text articles to assess study eligibility. The extracted data encompassed critical study characteristics, including the authors, year, country, study design, center, response definition, types of disorders, and types of treatments. Specific quantitative elements included the total number of patients in the training and internal validation sets, the number of response patients in the training and internal validation sets, the data-splitting method, predictors, AI algorithms, and AI methods. For internal validation sets, we systematically recorded true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

For studies included in the systematic review but lacking meta-analysis–compatible data, we proactively contacted corresponding authors via email to request necessary information. Given the frequent absence of diagnostic contingency tables, we reconstructed 2×2 tables using reported sensitivity, specificity, responder numbers, and total patient counts. Any discrepancies in extraction were resolved through a collaborative discussion, with a third reviewer (ZW) serving as an arbitrator if consensus could not be reached, thereby ensuring methodological rigor and data integrity.

Quality Assessment

The latest PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) quality assessment tool was used [24], which replaced PROBAST 2019. This tool comprises 2 phases: model development and model evaluation. Each phase encompasses 7 domains covering participant and data sources, predictors, outcome assessment, and analysis. Evaluation results for each domain are categorized as “low,” “high,” or “unclear,” determined based on specific signal questions. Signal concerns are rated as “yes,” “probably yes,” “probably no,” “no,” “no information,” and, in some cases, “not applicable.” Signal concerns phrased as “yes” or “probably yes” indicate a lower risk of bias. Any signal question rated “no” or “probably no” indicates a potential high risk of bias in that domain. If no “no” or “probably no” rating exists but “no information” is present, the domain’s bias risk is classified as unclear. The complete list of signal questions and tables is detailed in Tables S2 and S3 in Multimedia Appendix 1.

To ensure objectivity and accuracy in the assessment process, 2 reviewers (JL and JW) independently assessed the risk of bias in the included studies using the PROBAST+AI quality assessment tool. During the review process, any discrepancies between reviewers were resolved through in-depth discussions and analyses to achieve consensus, ensuring highly reliable, consistent final assessment results.

Outcome Measures

The primary outcome measures encompassed comprehensive diagnostic performance metrics from internal validation sets, including sensitivity, specificity, accuracy, and AUC. Sensitivity, calculated as TP/(TP+FN), measured the AI model's ability to accurately identify TPs, while specificity, computed as TN/(TN+FP), assessed the model's proficiency in correctly identifying negative cases. AUC provided a holistic assessment of the model's discriminative power between positive and negative instances. Accuracy, defined as (TP+TN)/total patients, represented the proportion of correctly classified samples. We systematically extracted performance metrics from all data sources (eg, imaging, clinical/demographic, and multimodal) and AI algorithms presented in each study. However, for the purpose of meta-analytic pooling, only the primary model from each article, specifically the model based on nonoverlapping patient cohorts, was selected to ensure data independence and avoid double-counting.

Statistical Analysis

While the core analytical plan and reporting adhered to the preregistered protocol, a key methodological deviation was implemented to enhance the robustness of our statistical synthesis. The registered protocol specified the use of the bivariate random-effects model (Reitsma model) as the primary method for pooling sensitivity and specificity. In the final analysis, we maintained this model for generating the summary receiver operating characteristic (SROC) curve and the AUC. However, for the pooled estimates of sensitivity and specificity themselves, we adopted a univariate random-effects model using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. This adaptation was made to incorporate the HKSJ adjustment, which provides more conservative and reliable CIs, particularly when dealing with a limited number of studies or substantial heterogeneity [25]. Prediction intervals (PIs) were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings. Heterogeneity was quantitatively assessed using the PIs and further explored through subgroup analyses and meta-regression [26]. Bivariate boxplots were used to visually identify potential outlier studies contributing to heterogeneity. Prespecified covariates for meta-regression included the number of study centers, anxiety disorder subtype, data splitting method, CBT protocol type, and AI algorithm category. Subgroup analyses were conducted according to data type, algorithm type, and specific anxiety disorder diagnoses. Additionally, a bubble plot was used to assess temporal trends in the accuracy of different algorithms, and funnel plots were used, along with Egger tests, to evaluate potential small-study effects.

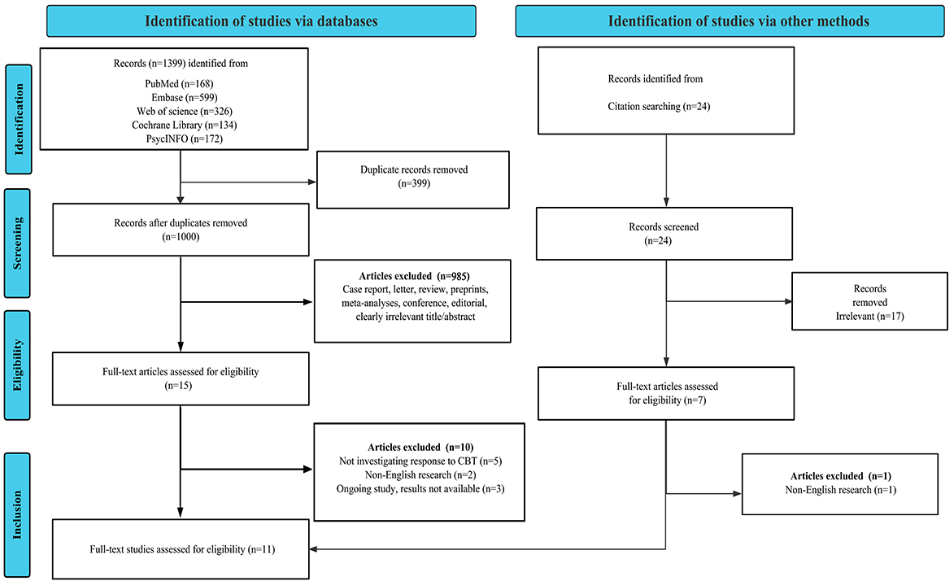
Results

Study Selection

A search across 5 electronic databases yielded 1399 potentially relevant publications. After deleting 399 duplicates, 1000 records were retained for preliminary screening. At this phase, 980 studies were identified as clearly irrelevant based on the title/abstract and the publication type. Thus, 15 articles were selected for full-text review. Following a comprehensive assessment of the full texts, 5 studies were removed because they did not investigate treatment response to CBT. Two studies were excluded due to non-English publication. Three studies were excluded as they were ongoing trials without any published outcome measures. Additionally, 6 qualifying studies were identified from nondatabase sources. Ultimately, 11 studies [7,17,18,27-34] were included in the meta-analysis as they met predefined inclusion criteria. The identification, screening, eligibility assessment, and inclusion process followed the PRISMA guidelines, and the selection analysis is elaborated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) s in Figure 1:

flow diagram of the systematic literature search and study selection process for studies evaluating artificial intelligence models in predicting treatment response to cognitive behavioral therapy for anxiety disorders.


Identification of studies via databases
Identification of studies via other methods
Identification
Screening
Eligibility
Inclusion
Records (n=1399) identified from
PubMed (n=168)
Embase (n=599)
Web of science (n=326)
Cochrane Library (n=134)
PsycINFO (n=172)
Records identified from
Citation searching (n=24)
Duplicate records removed (n=399)
Records after duplicates removed (n=1000)
Records screened (n=24)
Records removed Irrelevant (n=17)
Articles excluded (n=985)
Case report, letter, review, preprints, meta-analyses, conference, editorial, clearly irrelevant title/abstract
Full-text articles assessed for eligibility (n=15)
Full-text articles assessed for eligibility (n=7)
Articles excluded (n=10)
Not investigating response to CBT (n=5)
Non-English research (n=2)
Ongoing study, results not available (n=3)
Articles excluded (n=1)
Non-English research (n=1)
Full-text studies assessed for eligibility (n=11)

Study Description and Quality Assessment

A total of 11 eligible studies comprising 59,085 samples (training: 38,563; internal validation set: 20,522) were included. Among these, 4 studies exclusively evaluated SAD [7,29,32,34], 2 studies focused on PD with agoraphobia [30,33], 1 study investigated both GAD and PD [27], 1 study examined PD and SAD [31], and the remaining 3 studies addressed multiple anxiety disorder subtypes [17,18,28]. Nine studies [17,18,27,29-34] provided data eligible for meta-analysis. The included studies were published between 2014 and 2024, and all were prospective. Definitions of treatment response varied across studies: 2 studies used the CGI-I [29,32], 2 used the LSAS [7,34], 2 used the HAM-A [30,33], 1 used the Overall Anxiety Severity and Impairment Scale (OASIS) [27], 1 used the clinical severity rating [28], 1 combined HAM-A and the Spider Phobia Questionnaire (SPQ) [18], 1 used both the Generalized Anxiety Disorder-7 (GAD-7) and the Patient Health Questionnaire-9 (PHQ-9) [17], and 1 applied multiple scales simultaneously [31]. Across the 9 datasets, the most frequently used AI algorithms were support vector machines (SVMs; 4/9, 44%) and random forests (RFs; 3/9, 33%). Additionally, of the 9 datasets, 2 were based on clinical/demographic data, 6 on imaging data, and 1 on multimodal data. The clinical, methodological, and technical characteristics of the included studies are summarized in Table 1 and Multimedia Appendix 2.

Table 1. Characteristics of the included studies (n=11) and their patient populations.

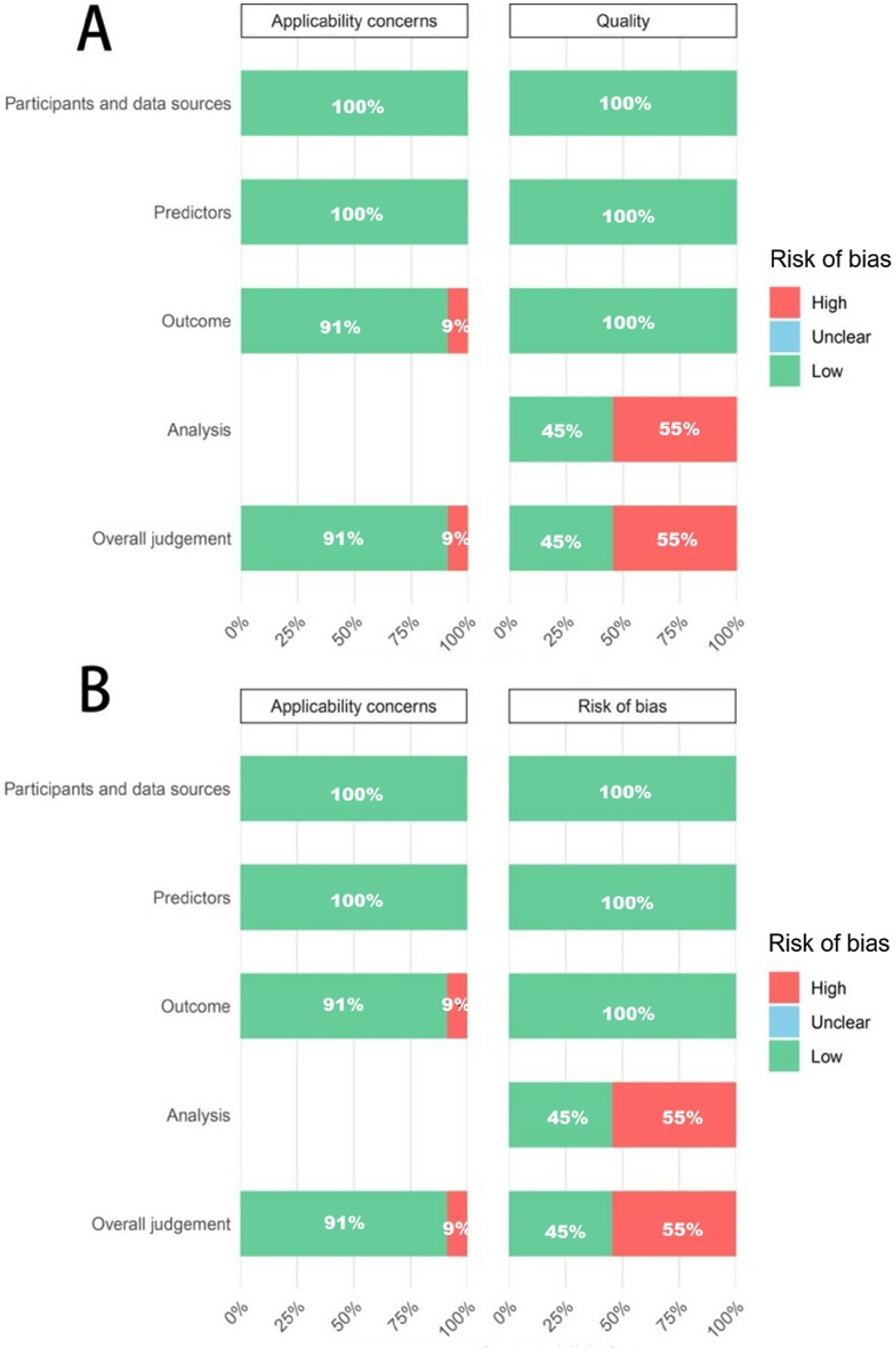
Study (author)	Year	Country	Study design	Center	Response definition	Types of anxiety disorders	Types of treatments	Data source	Total sample size, n		Response sample size, n	
									Training	Internal validation	Training	Internal validation
Ball et al [27]	2014	America	Pro[a]	Single center	OASIS[b] ≤5	GAD[c], PD[d]	Standard CBT[e]	Institutional database	48	49	29	49
Bertie et al [28]	2024	Multiple countries	Pro	Multiple centers	CSR[f] ≥4	GAD, separation anxiety, SAD[g], SPH[h]	Standard CBT	Institutional database	2214	2214	—[i]	—
Bukhari et al [7]	2025	America	Pro	Multiple centers	LSAS[j]↓ ≥50%	SAD	Standard CBT	Institutional database	157	157	—	—
Frick et al [29]	2020	Sweden	Pro	Single center	CGI-I[k] ≤2	SAD	CBT+SSRI[l],	Institutional database	47	47	24	24

Study (author)	Year	Country	Study design	Center	Response definition	Types of anxiety disorders	Types of treatments	Data source	Total sample size, n		Response sample size, n	
									Training	Internal validation	Training	Internal validation
							CBT+placebo					
Hahn et al [30]	2015	Germany	Pro	Multiple centers	HAM-A[m]↓ ≥50%	PD with AG[n]	Standard CBT	Institutional database	49	49	25	25
Hilbert et al [18]	2024	Germany	Pro	Multiple centers	Protect-AD: HAM-A↓ ≥50%; Spider VR: SPQ[o]↓ ≥30%	PD, AG, SAD, SPH	Standard CBT	Institutional database	410	410	213	213
Hentati Isacsson et al [31]	2024	Sweden	Pro	Single center	MADRS[p] ≤11, PDSS[q] ≤8, LSAS ≤35	PD, SAD	iCBT[r]	Institutional database	3619	3619	—	—
Månsson et al [32]	2015	Sweden	Pro	Multiple centers	CGI-I ≤2	SAD	iCBT	Institutional database	23	23	12	12
Prasad et al [17]	2023	United Kingdom, America	Pro	Multiple centers	GAD-7[s]↓ ≥64, PHQ-9[t]↑ <6	Anxiety	iCBT	Institutional database	31,899	13,857	—	5232
Sundermann et al [33]	2017	Germany	Pro	Multiple centers	HAM-A↓ ≥50%	PD with AG	Standard CBT	Institutional database	59	59	30	30
Whitfield-Gabrieli et al [34]	2015	America	Pro	Multiple centers	LSAS↓ ≥50%	SAD	Standard CBT	Institutional database	38	38	19	19

[a]Pro: prospective.
[b]OASIS: Overall Anxiety Severity and Impairment Scale.
[c]GAD: generalized anxiety disorder.
[d]PD: panic disorder.
[e]CBT: cognitive behavioral therapy.
[f]CSR: clinical severity rating.
[g]SAD: social anxiety disorder.
[h]SPH: specific phobia.
[i]Not applicable.
[j]LASA: Liebowitz Social Anxiety Scale.
[k]CGI-I: clinical global impression-improvement.
[l]SSRI: selective serotonin reuptake inhibitor.
[m]HAM-A: Hamilton Anxiety Rating Scale.
[n]AG: agoraphobia.
[o]SPQ: Spider Phobia Questionnaire.
[p]MADRS: Montgomery-Asberg Depression Rating Scale.
[q]PDSS: Panic Disorder Severity Scale.
[r]iCBT: internet-based cognitive behavioral therapy.
[s]GAD-7: Generalized Anxiety Disorder-7.
[t]PHQ-9: Patient Health Questionnaire-9.

The risk of bias, as assessed by the PROBAST+AI quality assessment tool, is presented in Figure 2 and Tables S2 and S3 in Multimedia Appendix 1. For model development, in the overall quality assessment, 55% (6/11) of studies were rated as high risk, while the remaining 45% (5/11) were rated as low risk. Regarding applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk. For model validation, 55% (6/11) of studies were rated as high risk in the overall risk of bias assessment, with the remaining 45% (5/11) rated as low risk. In the comprehensive evaluation of applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk.

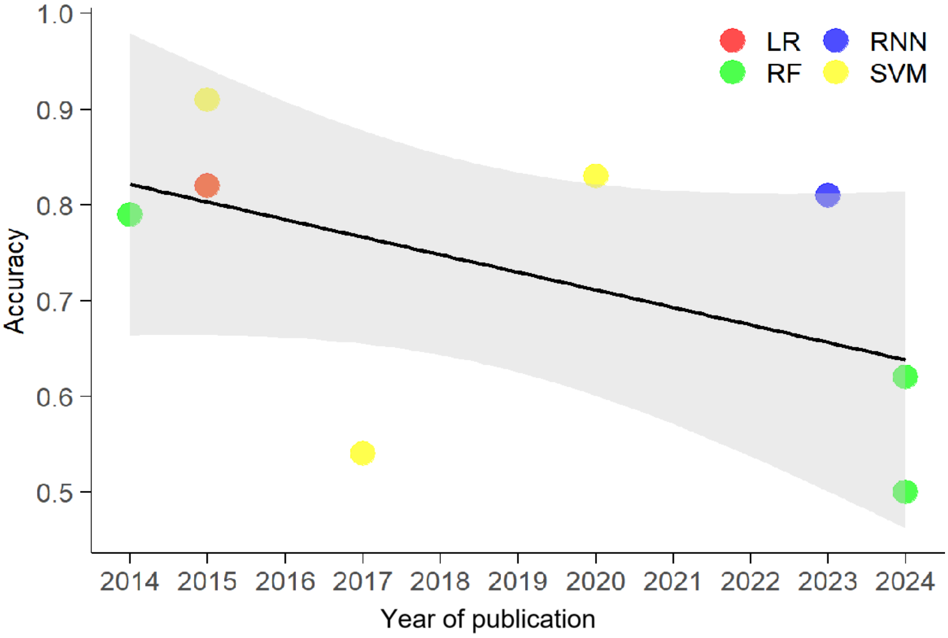
Figure 2. Risk of bias and applicability assessment using the PROBAST+AI tool for studies developing or validating AI models to predict treatment response after cognitive behavioral therapy in patients with anxiety disorders. Panel A summarizes the risk of bias for model development studies, and panel B summarizes the risk of bias for model validation studies. PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence.



Trend in Diagnostic Accuracy Over Time

The scatter plot in Figure 3 illustrates a negative trend in model accuracy from 2014 to 2024 across logistic regression (LR), recurrent neural network (RNN), RF, and SVM algorithms. A linear regression line confirms that reported accuracy has generally decreased over the decade within this dataset. The highest performance was recorded in 2015 by an SVM model (approximately 0.9), while the lowest scores appeared in 2024, with RF models showing a drop in scores to around 0.5. The shaded region indicates the CI for this downward trajectory (Figure 3).

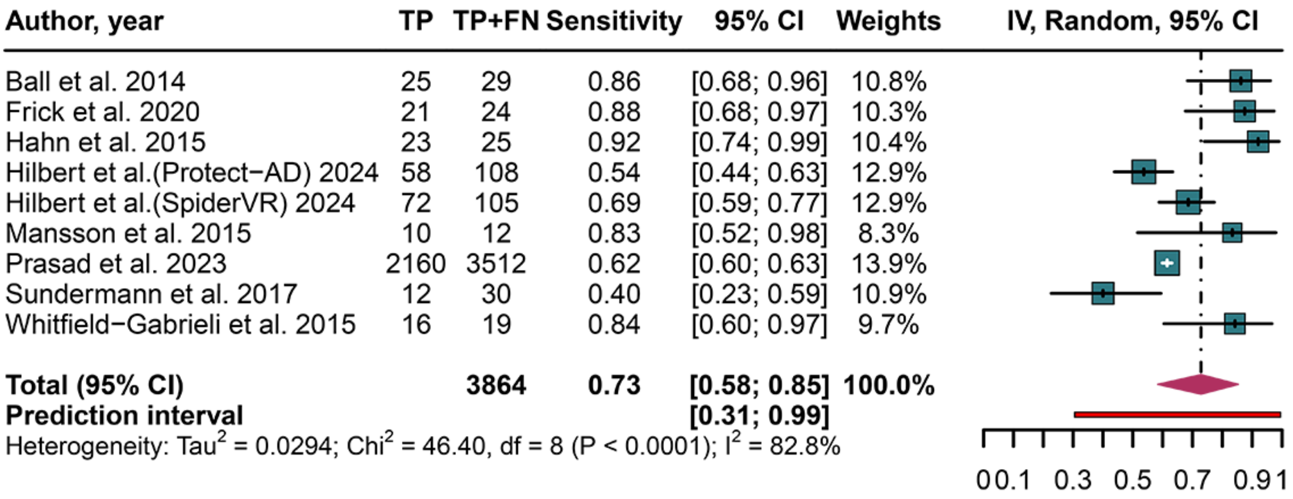
Figure 3. Bubble plot illustrating the accuracy of various artificial intelligence algorithms over time (publication year) for predicting treatment response to cognitive behavioral therapy in patients with anxiety disorders. LR: logistic regression; RF: random forest; RNN: recurrent neural network; SVM: support vector machine.



Predicting CBT Response to Anxiety Disorders Using AI Models Based on the Primary Model Per Study

The pooled sensitivity of AI models was 0.73 (95% CI 0.58-0.85; 95% PI 0.31-0.99; I^2=82.8%), specificity was 0.75 (95% CI 0.59-0.89; 95% PI 0.25-1.00; I^2=96.7%), accuracy was 0.74 (95% CI 0.62-0.84; 95% PI 0.39-0.97; I^2=94.6%), and AUC was 0.81 (95% CI 0.78-0.85), as shown in Figures 4-7. Based on a prespecified pretest probability of 20%, the Fagan nomogram indicated a posttest probability of 43% for a positive result and 8% for a negative result (Figure S1 in Multimedia Appendix 1). The results of the bivariate boxplots indicated that 1 study [30] had discrete values, suggesting it may be a source of heterogeneity (Figure 8). Furthermore, meta-regression analysis indicated that heterogeneity in sensitivity was significantly associated with the number of centers (*P*<.001) and the type of anxiety disorder (*P*<.001). In contrast, heterogeneity in specificity was significantly associated with CBT type (*P*<.001) and AI algorithm category (*P*<.001) (Table 2).

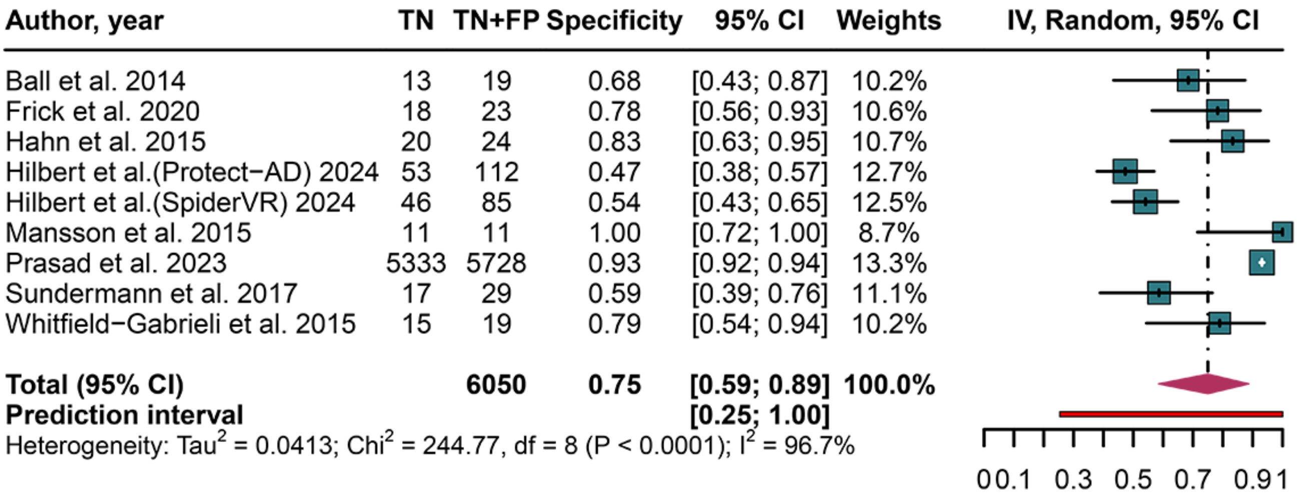
Figure 4. Forest plot displaying the pooled sensitivity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FN: false negative; TP: true positive.



Author, year	TP	TP+FN	Sensitivity	95% CI	Weights
Ball et al. 2014	25	29	0.86	[0.68; 0.96]	10.8%
Frick et al. 2020	21	24	0.88	[0.68; 0.97]	10.3%
Hahn et al. 2015	23	25	0.92	[0.74; 0.99]	10.4%
Hilbert et al.(Protect−AD) 2024	58	108	0.54	[0.44; 0.63]	12.9%
Hilbert et al.(SpiderVR) 2024	72	105	0.69	[0.59; 0.77]	12.9%
Mansson et al. 2015	10	12	0.83	[0.52; 0.98]	8.3%
Prasad et al. 2023	2160	3512	0.62	[0.60; 0.63]	13.9%
Sundermann et al. 2017	12	30	0.40	[0.23; 0.59]	10.9%
Whitfield−Gabrieli et al. 2015	16	19	0.84	[0.60; 0.97]	9.7%
Total (95% CI)		**3864**	**0.73**	**[0.58; 0.85]**	**100.0%**
Prediction interval				**[0.31; 0.99]**	

Heterogeneity: Tau^2 = 0.0294; Chi^2 = 46.40, df = 8 (P < 0.0001); I^2 = 82.8%

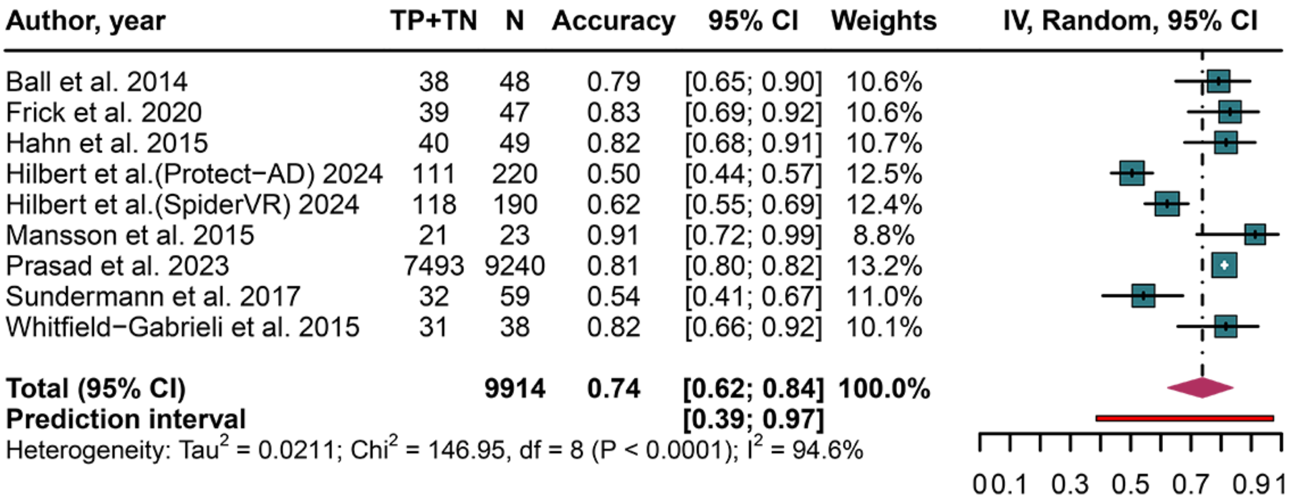
Figure 5. Forest plot displaying the pooled specificity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FP: false positive; TN: true negative.



Author, year	TN	TN+FP	Specificity	95% CI	Weights
Ball et al. 2014	13	19	0.68	[0.43; 0.87]	10.2%
Frick et al. 2020	18	23	0.78	[0.56; 0.93]	10.6%
Hahn et al. 2015	20	24	0.83	[0.63; 0.95]	10.7%
Hilbert et al.(Protect−AD) 2024	53	112	0.47	[0.38; 0.57]	12.7%
Hilbert et al.(SpiderVR) 2024	46	85	0.54	[0.43; 0.65]	12.5%
Mansson et al. 2015	11	11	1.00	[0.72; 1.00]	8.7%
Prasad et al. 2023	5333	5728	0.93	[0.92; 0.94]	13.3%
Sundermann et al. 2017	17	29	0.59	[0.39; 0.76]	11.1%
Whitfield−Gabrieli et al. 2015	15	19	0.79	[0.54; 0.94]	10.2%
Total (95% CI)		**6050**	**0.75**	**[0.59; 0.89]**	**100.0%**
Prediction interval				**[0.25; 1.00]**	

Heterogeneity: Tau^2 = 0.0413; Chi^2 = 244.77, df = 8 (P < 0.0001); I^2 = 96.7%

Figure 6. Forest plot displaying the pooled accuracy, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. TN: true negative; TP: true positive.



Author, year	TP+TN	N	Accuracy	95% CI	Weights
Ball et al. 2014	38	48	0.79	[0.65; 0.90]	10.6%
Frick et al. 2020	39	47	0.83	[0.69; 0.92]	10.6%
Hahn et al. 2015	40	49	0.82	[0.68; 0.91]	10.7%
Hilbert et al.(Protect−AD) 2024	111	220	0.50	[0.44; 0.57]	12.5%
Hilbert et al.(SpiderVR) 2024	118	190	0.62	[0.55; 0.69]	12.4%
Mansson et al. 2015	21	23	0.91	[0.72; 0.99]	8.8%
Prasad et al. 2023	7493	9240	0.81	[0.80; 0.82]	13.2%
Sundermann et al. 2017	32	59	0.54	[0.41; 0.67]	11.0%
Whitfield−Gabrieli et al. 2015	31	38	0.82	[0.66; 0.92]	10.1%
Total (95% CI)		**9914**	**0.74**	**[0.62; 0.84]**	**100.0%**
Prediction interval				**[0.39; 0.97]**	

Heterogeneity: Tau^2 = 0.0211; Chi^2 = 146.95, df = 8 (P < 0.0001); I^2 = 94.6%

Figure 7. Summary receiver operating characteristic (SROC) curve for artificial intelligence models predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. The summary point estimates for sensitivity (SENS) and specificity (SPEC), and the area under the curve (AUC) are shown.

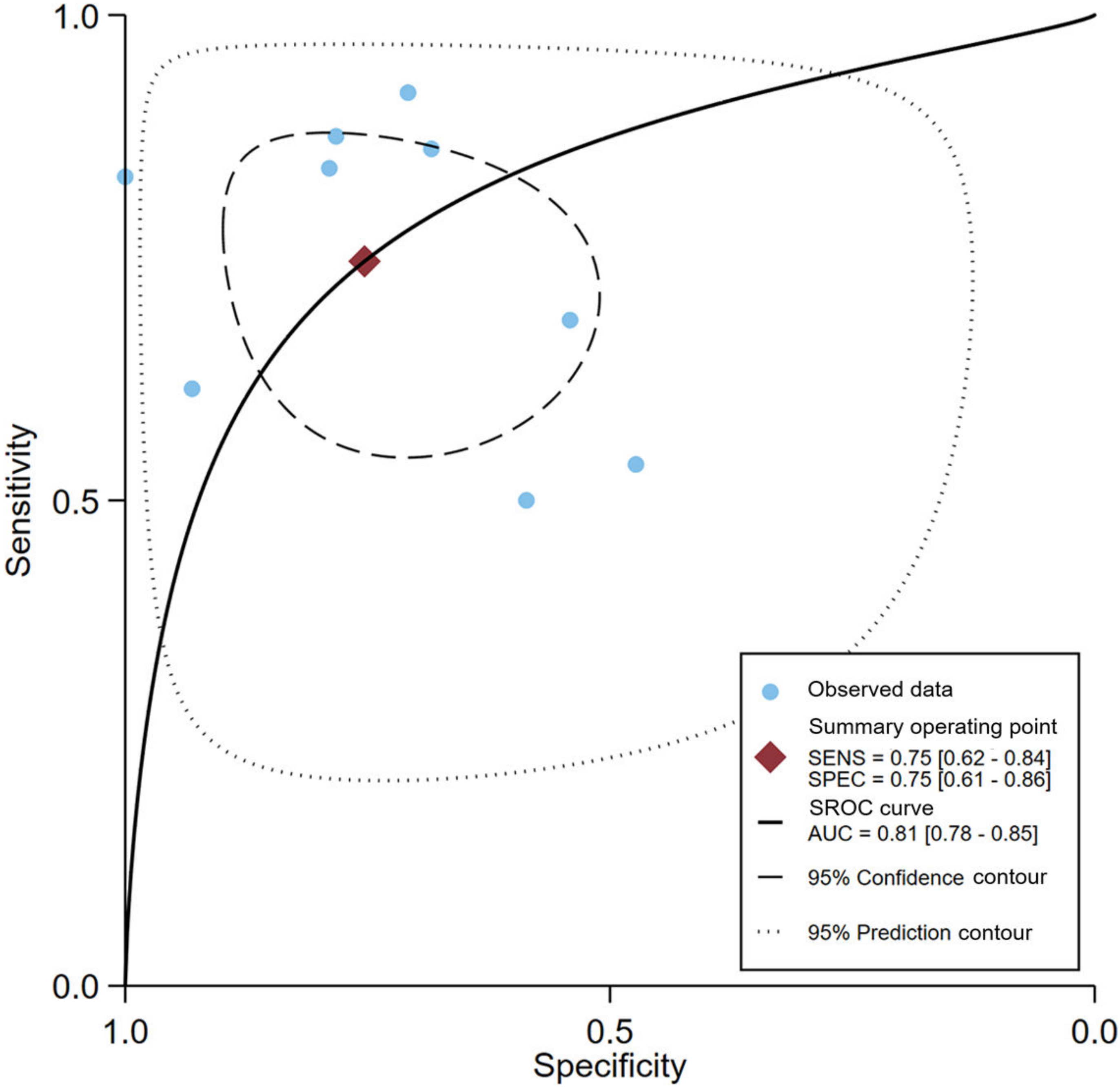


Figure 8. Bivariate boxplot of the combined logit-transformed sensitivity (SENS) and specificity (SPEC) estimates from the bivariate random-effects meta-analysis for predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. Squares represent individual study estimates; the box illustrates the joint distribution.

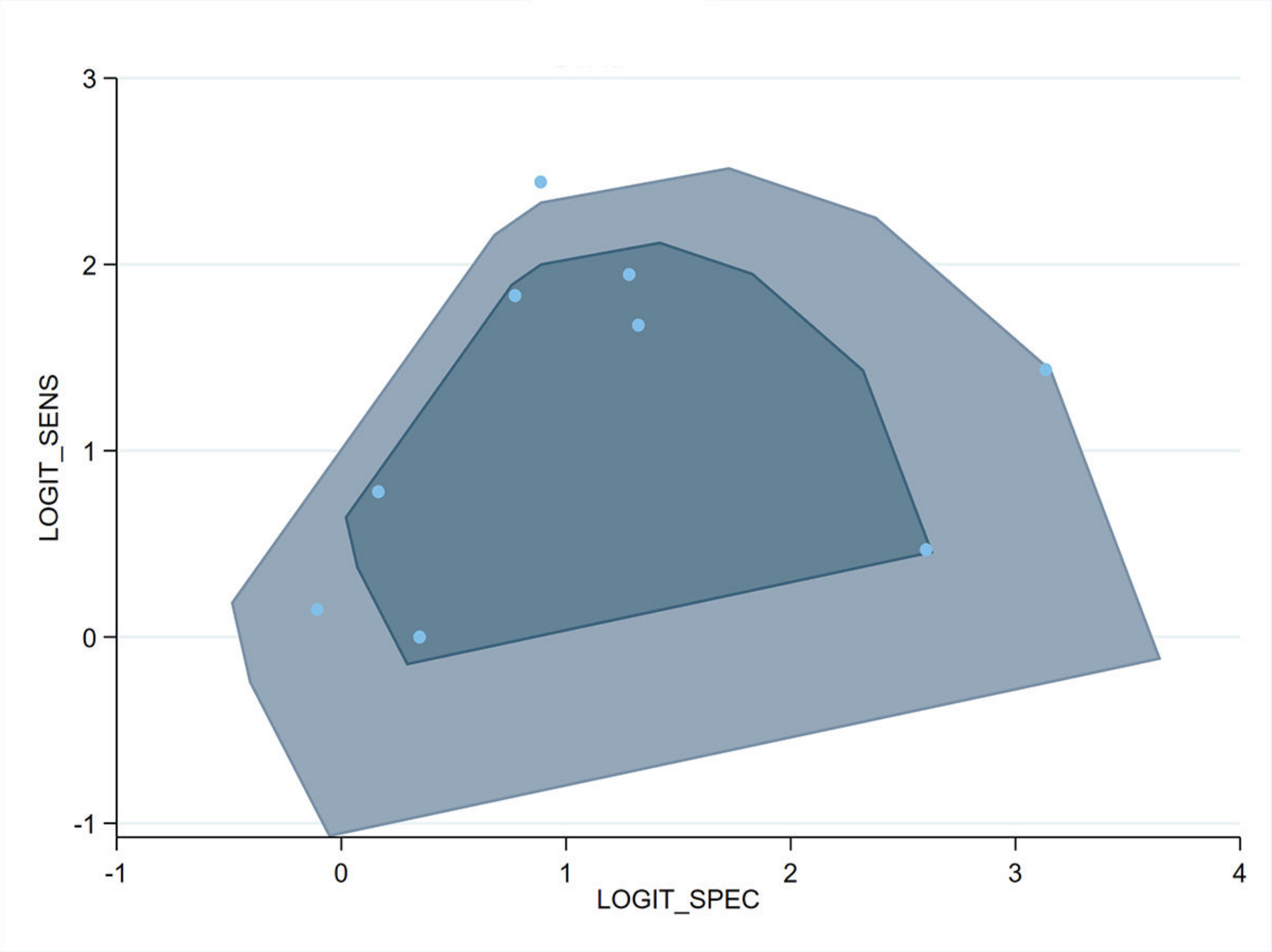


Table 2. Results of the univariable meta-regression analysis examining potential sources of heterogeneity in the performance of artificial intelligence models for predicting cognitive behavioral therapy treatment response in anxiety disorders.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Meta-regression *P* value	Specificity, value (95% CI)	Meta-regression *P* value
Number of centers			.02		.89
Single center	2	0.87 (0.76-0.99)		0.74 (0.46-1.00)	
Multiple centers	7	0.69 (0.58-0.80)		0.76 (0.61-0.90)	
Type of anxiety disorder			.01		.10
SAD[a]	3	0.86 (0.74-0.98)		0.87 (0.71-1.00)	
Others	6	0.68 (0.57-0.79)		0.69 (0.54-0.85)	
Data splitting method			.23		.28
LOOCV[b]	4	0.78 (0.62-0.93)		0.81 (0.63-0.98)	
Independent validation	5	0.72 (0.58-0.86)		0.71 (0.54-0.89)	
CBT[c] type			.77		<.001
iCBT[d]	2	0.71 (0.48-0.94)		0.95 (0.91-0.98)	
Standard CBT	6	0.74 (0.62-0.87)		0.61 (0.51-0.71)	
AI[e] algorithms			.51		<.001
DL[f]	1	0.62 (0.29-0.94)		0.93 (0.87-1.00)	
ML[g]	8	0.78 (0.68-0.88)		0.67 (0.57-0.78)	

[a]SAD: social anxiety disorder.

[b]LOOCV: leave-one-out cross-validation.
[c]CBT: cognitive behavioral therapy.
[d]iCBT: internet-based cognitive behavioral therapy.
[e]AI: artificial intelligence.
[f]DL: deep learning.
[g]ML: machine learning.

Subgroup Analysis

Within the data-type subgroup, multimodal AI models appeared to achieve higher pooled sensitivity (0.84, 95% CI 0.60-0.97) and accuracy (0.82, 95% CI 0.66-0.92) compared with single-modal approaches. In contrast, models based on either clinical/demographic or imaging data alone demonstrated comparatively lower sensitivity and accuracy. Regarding specificity, models developed using clinical/demographic data showed relatively higher estimates (0.81, 95% CI 0.55-0.93). Regarding AUC, only 1 study reported data within this subgroup, and in that study, imaging models achieved an AUC of 0.79 (95% CI 0.76-0.83) (Table 3).

Table 3. Subgroup meta-analysis of the pooled predictive performance of artificial intelligence models for predicting cognitive behavioral therapy response, stratified by input data type, algorithm type, and anxiety disorder classification.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Specificity, value (95% CI)	AUC[a], value (95% CI)	Accuracy, value (95% CI)
Type of data					
Clinical/demographic	2	0.65 (0.42-0.82)	0.81 (0.55-0.93)	—[b]	0.73 (0.55-0.89)
Imaging	6	0.77 (0.63-0.87)	0.71 (0.53-0.85)	0.79 (0.76-0.83)	0.72 (0.00-1.00)
Multimodal	1	0.84 (0.60-0.97)	0.79 (0.54-0.94)	—	0.82 (0.66-0.92)
AI[c] algorithms					
RF[d]	3	0.64 (0.58-0.70)	0.52 (0.45-0.59)	0.54 (0.50-0.59)	0.63 (0.27-0.93)
SVM[e]	4	0.82 (0.61-0.93)	0.75 (0.60-0.85)	0.83 (0.80-0.86)	0.78 (0.50-0.97)
RNN[f]	1	0.62 (0.60-0.63)	0.93 (0.92-0.94)	—	0.81 (0.81-0.82)
LR[g]	1	0.84 (0.60-0.97)	0.79 (0.54-0.94)	—	0.82 (0.66-0.92)
Type of anxiety disorder					
GAD[h] and PD[i]	1	0.86 (0.68-0.96)	0.68 (0.43-0.87)	—	0.79 (0.65-0.90)
SAD[j]	3	0.86 (0.78-0.92)	0.87 (0.41-1.00)	0.92 (0.85-0.98)	0.85 (0.72-0.94)
PD with AG[k]	2	0.69 (0.00-1.00)	0.71 (0.00-1.00)	—	0.69 (0.00-1.00)
Multiple anxiety disorders	3	0.61 (0.45-0.77)	0.68 (0.05-1.00)	0.65 (0.49-0.81)	0.65 (0.25-0.96)

[a]AUC: area under curve.
[b]Not applicable.
[c]AI: artificial intelligence.
[d]RF: random forest.
[e]SVM: support vector machine.
[f]RNN: recurrent neural network.
[g]LR: logistic regression.
[h]GAD: generalized anxiety disorder.
[i]PD: panic disorder.
[j]SAD: social anxiety disorder.
[k]AG: agoraphobia.

Within the AI algorithm subgroup, different algorithms demonstrated distinct advantages across various performance metrics. For sensitivity, LR performed best, achieving a value of 0.84 (95% CI 0.60-0.97). Similarly, in terms of accuracy, LR again showed optimal performance, reaching a value of 0.82 (95% CI 0.66-0.92). For specificity, the RNN achieved the highest value of 0.93 (95% CI 0.87-0.98). Regarding the AUC, only 2 algorithms produced results that could be combined, with the SVM algorithm yielding a relatively high combined AUC of 0.83 (95% CI 0.80-0.86).

Within the type of anxiety disorder subgroup, predictive performance varied across outcome metrics. For sensitivity, GAD and PD demonstrated relatively higher pooled estimates at 0.86 (95% CI 0.68-0.96). In contrast, SAD showed comparatively stronger performance in specificity and accuracy, with a pooled specificity of 0.87 (95% CI 0.41-1.00) and accuracy of 0.85 (95% CI 0.72-0.94). For discrimination, as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85-0.98).

Publication Bias

Funnel plots and Egger regression tests were conducted to evaluate potential small-study effects (Figures S2-S4 in

Multimedia Appendix 1). No statistically significant evidence of small-study effects was observed for sensitivity (t_7=1.75; *P*=.12; intercept=1.54; SE=0.88) and accuracy (t_7=−1.51; *P*=.17; intercept=−2.40; SE=1.59). In contrast, the test for specificity yielded a statistically significant result (t_7=−2.58; *P*=.04; intercept=−4.29; SE=1.66), suggesting the possible presence of small-study effects.

Discussion

Principal Findings

The systematic review and meta-analysis revealed that the model demonstrated moderate predictive performance in forecasting treatment response to CBT among patients with anxiety disorders, with sensitivity, specificity, accuracy, and AUC values of 0.73, 0.75, 0.74, and 0.81, respectively. This moderate predictive capability reflects the combined influence of multiple complex factors. First, limitations in input variables constitute a key constraint: current models primarily rely on demographic data, psychiatric history, and scale measurements, making it challenging to capture deeper biological and environmental factors [7,35]. While biomarkers, such as neuroimaging, could theoretically significantly enhance predictive accuracy, their clinical application remains constrained by cost and accessibility [32,35]. Second, data heterogeneity and sample size issues cannot be overlooked. Significant variations among patient cohorts in symptom severity, comorbid conditions, and treatment modalities diminish the model's generalizability [18,36]. Small-sample studies often overestimate predictive capability, while large-scale multicenter studies frequently yield reduced accuracy, sometimes approaching random levels [18,35].

Furthermore, limitations exist in variable selection and modeling approaches: prediction capabilities relying solely on traditional questionnaires and demographic data are constrained [35,37]. While integrating multimodal data (eg, neuroimaging, genetic, and behavioral data) may enhance performance, the comprehensive collection of such data remains challenging in real-world clinical settings [18]. Finally, challenges in practical application and external validation cannot be overlooked; many models lack independent external validation, resulting in significantly reduced accuracy and interpretability in clinical settings [38]. In summary, while current models demonstrate predictive potential, future research requires larger-scale, multicenter, and multimodal studies to enhance model robustness and clinical translation capabilities.

The subgroup analysis suggested that multimodal data integration was associated with comparatively higher sensitivity (0.84) and accuracy (0.82) than single-modal approaches. Across algorithm categories, no single model consistently outperformed others; LR showed higher sensitivity (0.84) and accuracy (0.82), RNN demonstrated higher specificity (0.93), and SVM had a comparatively higher AUC of 0.83. Predictive performance also varied across anxiety disorder subtypes. GAD and PD showed comparatively higher sensitivity (0.86), whereas SAD demonstrated higher specificity (0.87), accuracy (0.85), and AUC (0.92). Together, these findings suggest that optimal prediction may depend less on a single "best" algorithm and more on strategically matching the model architecture to the data type and clinical question [39]. Future work should therefore focus not only on multimodal data collection but also on developing tailored, interpretable modeling approaches that can be validated in real-world clinical settings [40].

The systematic review revealed that multimodal data demonstrated the most promising performance in predicting CBT treatment response for anxiety disorders, with a sensitivity of 0.84 and an accuracy of 0.82. The superior predictive capability of multimodal approaches could be attributed to several key mechanisms. First, multimodal data integration enabled the comprehensive capture of the multidimensional nature of anxiety disorders by combining neuroimaging, physiological signals, behavioral data, and clinical assessments [29,30]. Complementary information across modalities enabled AI models to detect nuanced individual differences and potential mechanisms of treatment response [27,34]. Research indicates that, compared with single-modality approaches, multimodal fusion significantly enhances predictive accuracy, particularly when incorporating models that integrate functional and structural brain connectivity [29,34]. Moreover, multimodal approaches supported personalized, dynamic prediction models that can identify high-risk nonresponders early, potentially facilitating adaptive treatment strategies [17,27]. While these findings are promising, careful validation through larger, cross-validated studies is essential to establish the clinical utility and generalizability of such AI-driven prediction models [29,33].

The meta-analysis revealed differential performance patterns across ML algorithms in predicting CBT treatment response for anxiety disorders. LR demonstrated comparatively higher sensitivity (0.84) and accuracy (0.82), potentially attributable to its ability to capture linear relationships between clinical features and treatment outcomes [7,31]. The algorithm's strength lies in its ability to model structured clinical data with minimal risk of overfitting, particularly in scenarios with balanced sample distributions [7,17]. Conversely, RNNs exhibited superior specificity (0.93), likely due to their ability to process high-dimensional, sequential data and capture complex nonlinear interactions [17]. RNN's advanced modeling approach enables more nuanced identification of nonresponders by leveraging temporal patterns in clinical measurements and platform interaction behaviors [17]. For overall discrimination, SVM achieved a pooled AUC of 0.83 (95% CI 0.80-0.86). Given their kernel-based formulation, SVMs may perform well with moderate sample sizes and complex feature spaces, allowing flexible decision boundaries [7,31]. Taken together, these findings suggest that algorithmic performance may vary according to data structure and modeling objectives. Rather than indicating a universally superior approach, the results support the importance of aligning model architecture with data characteristics and specific clinical prediction goals [31].

The subgroup analysis indicated that predictive performance varied across anxiety disorder subtypes. GAD and PD demonstrated comparatively higher sensitivity estimates (0.86, 95% CI 0.68-0.96), suggesting relatively stronger identification of treatment responders in these populations. In contrast, SAD showed higher pooled specificity (0.87, 95% CI 0.41-1.00) and accuracy (0.85, 95% CI 0.72-0.94). For discrimination as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85-0.98). The variation in predictive performance across anxiety disorder subtypes may partly account for the observed heterogeneity. In the subgroup analysis, GAD and PD demonstrated higher sensitivity, which may be related to the use of structured symptom-based scales that provide relatively stable responder signals for model training [27]. First, SAD research often relies on symptom measures that closely align core fear and avoidance dimensions with treatment outcome definitions, which may facilitate clearer model discrimination [7,34]. Second, some SAD studies incorporate high-information features beyond routine clinical variables, potentially enhancing overall predictive performance [29,32]. Third, SAD samples are frequently drawn from relatively homogeneous clinical trials with standardized intervention protocols, which may reduce heterogeneity and improve model stability [7,32].

Comparison With Prior Work

In the systematic review and meta-analysis published by Vieira et al [35] in 2022, the authors evaluated ML approaches for predicting CBT treatment response across multiple psychiatric disorders. Their analysis included 24 studies comprising 7497 patients and covered 5 diagnostic groups: major depressive disorder, obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), anxiety disorders, and substance use disorders. The overall pooled accuracy was 0.74 (95% CI 0.70-0.78), with variability across diagnostic categories. PTSD, anxiety disorders, and OCD demonstrated comparatively higher accuracy estimates (0.787, 0.776, and 0.761, respectively). In comparison, our study focused specifically on anxiety disorders and yielded a pooled accuracy of 0.74 (95% CI 0.62-0.84), which is comparable to the overall estimate reported by Vieira et al [35]. However, our analysis extended beyond accuracy by systematically synthesizing sensitivity, specificity, and AUC to provide a more comprehensive assessment of predictive performance. In addition, we conducted subgroup analyses examining the potential influence of data types (eg, clinical and neuroimaging), algorithm types, and disease types on model performance, thereby offering a more granular evaluation within the anxiety disorder domain. Bubble plots were used to illustrate temporal trends and the distribution of study characteristics across performance metrics.

Heterogeneity

The substantial heterogeneity observed across studies may influence the overall strength of evidence for AI-based prediction models. Given the anticipated between-study variability, a bivariate random-effects model was applied to pool sensitivity and specificity estimates. Meta-regression and subgroup analyses were further conducted to explore potential sources of heterogeneity. The results indicated that heterogeneity in sensitivity was significantly associated with the number of study centers (*P*=.02) and the anxiety disorder subtype (*P*=.01). In contrast, heterogeneity in specificity was significantly associated with CBT type (*P*<.001) and AI algorithm category (*P*<.001). The comparatively higher sensitivity observed in single-center studies, compared with multicenter studies, may reflect more homogeneous samples and more consistent data collection procedures [41]. Differences across anxiety disorder types may relate to variations in symptom structure, measurement alignment, and sample characteristics, which could influence the clarity of the responder signals available to the models [42]. Meanwhile, differences between iCBT and standard CBT may reflect variations in intervention structure and in the intensity of monitoring [43]. Variability in specificity across algorithm categories (DL vs ML) may relate to differences in model complexity and data representation strategies [44].

The bivariate boxplot suggested that 1 study [30] may represent a potential source of heterogeneity. Nevertheless, heterogeneity is likely multifactorial. Additional contributors may include differences in patient age, disorder severity, prior treatment exposure, geographic setting, sample size, feature selection strategies, data preprocessing methods, and hyperparameter optimization [45-48]. The combined influence of these methodological and clinical factors may account for the variability in predictive performance. Future research would benefit from more standardized reporting and systematic consideration of these variables to enhance reproducibility and external validity.

Implications for Practice and Research

Discussing the clinical interpretation of our results, the AI models demonstrated moderate predictive value for CBT treatment response in anxiety disorders, with the strongest performance observed for SAD. Multimodal models, which combine clinical demographics with imaging data, showed relatively superior predictive performance compared with unimodal approaches. This suggests that AI could support clinicians in making earlier, more informed therapeutic decisions and potentially improve patient outcomes by enabling timely treatment adjustments [49,50]. Specifically, these predictive models could be integrated into clinical workflows at key decision points: (1) during initial assessment to help stratify patients by likelihood of response, informing the intensity of early monitoring or the consideration of augmented treatment protocols, and (2) during the early stages of CBT to identify potential nonresponders, allowing for a timely switch or integration of other evidence-based interventions before disengagement occurs [35,51].

However, AI should serve as a decision-support tool rather than replacing clinician judgment, since treatment choices for anxiety disorders are influenced by not only severity but also patient preferences, comorbidity, and functional goals [52]. In the treatment of anxiety disorders, complex therapeutic decisions not only depend on assessing disease severity but also closely relate to the patient's individual condition and needs [53]. For real-world deployment, developing AI tools as clinician-facing dashboards that present predictive probabilities alongside key clinical context, rather than as autonomous decision systems, is crucial for fostering trust and facilitating shared decision-making [54]. Notably, among the studies we included, external validation sets were relatively scarce, and no research has yet focused solely on predicting outcomes for external validation cohorts. Future studies are needed to evaluate the generalization capabilities of AI models. Furthermore, limitations to the future adoption of AI include scarce annotated anxiety disorder data and regulatory hurdles [55]. Technical challenges persist in data availability, model interpretability, and transparency [56]. Advances in few-shot learning, self-supervised models, and centralized platforms may support an integrated AI ecosystem [57].

Limitations

When interpreting the findings of this systematic review and meta-analysis on the AI prediction of CBT treatment response for anxiety disorders, several limitations should be considered. First, the PROBAST+AI assessment indicated a high risk of bias in over half of the included studies, particularly in analytical methodology, which may have influenced the pooled estimates and contributed to between-study heterogeneity. Future studies should adopt more rigorous design and transparent reporting standards. Second, variability in the definition of treatment response across studies may have affected comparability. Although most scales assessed anxiety symptom reduction, differences in operational criteria could influence predictive performance. Greater standardization of outcome definitions would enhance consistency in future research. Third, the substantial heterogeneity and wide PIs suggest that model performance may vary considerably across clinical settings. Therefore, the pooled estimates should be interpreted cautiously and regarded as preliminary within the context of current methodological constraints. Fourth, although heterogeneity was observed across studies, this likely reflects differences in study design, patient characteristics, and modeling approaches. The reported PIs indicate that performance may vary across settings, highlighting the importance of external validation in future research. Finally, due to limited data availability, direct head-to-head comparisons between AI models and routinely used clinical scales were not feasible. Future research should incorporate comparative designs to better assess the incremental value of AI-based prediction tools. Overall, while AI models demonstrate potential for supporting treatment decision-making, stronger evidence from externally validated and methodologically robust studies is needed before clinical implementation.

Conclusions

This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying SAD as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation, and reformatting. The GAI tool used was Sider. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by JL.

Funding

The authors declare that no funding was received for this study. No funding body had any role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL contributed to conceptualization, critical revision of the manuscript, and final approval. JL, JW, ZW, and MISBAA contributed to data curation, formal analysis, investigation, and methodology. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Data to support the study.
[DOCX File (Microsoft Word File), 253 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Technical aspects and diagnostic performance data extracted from the included studies.
[DOCX File (Microsoft Word File), 27 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-DTA checklist.
[DOCX File (Microsoft Word File), 22 KB-Checklist 1]

Checklist 2

PRISMA 2020 Abstract checklist
[DOCX File (Microsoft Word File), 20 KB-Checklist 2]

Checklist 3

PRISMA-S checklist
[DOCX File (Microsoft Word File), 20 KB-Checklist 3]

References

1. Chodavadia P, Teo I, Poremski D, Fung DSS, Finkelstein EA. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: results from a 2022 web panel. BMC Psychiatry. Feb 14, 2023;23(1):104. [doi: 10.1186/s12888-023-04581-7] [Medline: 36782116]
2. Keramat SA, Comans T, Pearce A, Basri R, Hashmi R, Dissanayaka NN. Psychological distress and productivity loss: a longitudinal analysis of Australian working adults. Eur J Health Econ. Nov 2025;26(8):1503-1524. [doi: 10.1007/s10198-025-01764-9] [Medline: 40304834]
3. Powell C, Chiu CY, Sun X, So SHW. A meta-analysis on the efficacy of low-intensity cognitive behavioural therapy for generalised anxiety disorder. BMC Psychiatry. Jan 2, 2024;24(1):10. [doi: 10.1186/s12888-023-05306-6] [Medline: 38166836]
4. Papola D, Miguel C, Mazzaglia M, et al. Psychotherapies for generalized anxiety disorder in adults: a systematic review and network meta-analysis of randomized clinical trials. JAMA Psychiatry. Mar 1, 2024;81(3):250-259. [doi: 10.1001/jamapsychiatry.2023.3971] [Medline: 37851421]
5. Kravchenko O, Bäckman J, Mataix-Cols D, et al. Clinical, genetic, and sociodemographic predictors of symptom severity after internet-delivered cognitive behavioural therapy for depression and anxiety. BMC Psychiatry. May 30, 2025;25(1):555. [doi: 10.1186/s12888-025-07012-x] [Medline: 40448103]
6. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med. Sep 15, 2022;5(1):144. [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
7. Bukhari Q, Rosenfield D, Hofmann SG, Gabrieli JDE, Ghosh SS. Predicting treatment response to cognitive behavior therapy in social anxiety disorder on the basis of demographics, psychiatric history, and scales: a machine learning approach. PLoS ONE. 2025;20(3):e0313351. [doi: 10.1371/journal.pone.0313351] [Medline: 40100879]
8. Villarreal-Zegarra D, Paredes-Angeles R, Mayo-Puchoc N, Arenas-Minaya E, Huarcaya-Victoria J, Copez-Lonzoy A. Psychometric properties of the GAD-7 (General Anxiety Disorder-7): a cross-sectional study of the Peruvian general population. BMC Psychol. Apr 2, 2024;12(1):183. [doi: 10.1186/s40359-024-01688-8] [Medline: 38566138]
9. Manzar MD, Kashoo FZ, Salahuddin M, et al. Hamilton Rating Scale for Anxiety: exploring validity with robust measures of classical theory parameters and a rating scale model in university students. BJPsych Open. Aug 12, 2025;11(5):e176. [doi: 10.1192/bjo.2025.10055] [Medline: 40790884]
10. Zhang CY, Voort JLV, Yuruk D, et al. A characterization of the Clinical Global Impression Scale thresholds in the treatment of adolescent depression across multiple rating scales. J Child Adolesc Psychopharmacol. Jun 2022;32(5):278-287. [doi: 10.1089/cap.2021.0111] [Medline: 35704877]
11. Arrow K, Resnik P, Michel H, et al. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. Psychiatr Q. Jun 2023;94(2):221-231. [doi: 10.1007/s11126-023-10022-1] [Medline: 37145257]
12. Wang S, Lu J, Zheng G, et al. Poor performance of PHQ-9 and GAD-7 in screening clinical depression and anxiety among a large sample of Chinese children and adolescents. BMC Psychiatry. Mar 29, 2025;25(1):301. [doi: 10.1186/s12888-025-06754-y] [Medline: 40158158]

13. Bertie LA, Arendt K, Coleman JRI, et al. Patterns of sub-optimal change following CBT for childhood anxiety. J Child Psychol Psychiatry. Dec 2024;65(12):1612-1623. [doi: 10.1111/jcpp.14009] [Medline: 38817012]
14. Park JH, Shin YB, Jung D, et al. Machine learning prediction of anxiety symptoms in social anxiety disorder: utilizing multimodal data from virtual reality sessions. Front Psychiatry. 2024;15:1504190. [doi: 10.3389/fpsyt.2024.1504190] [Medline: 39896993]
15. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res. Jul 10, 2023;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
16. Habicht J, Dina LM, McFadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J Med Internet Res. Mar 10, 2025;27:e60435. [doi: 10.2196/60435] [Medline: 40063074]
17. Prasad N, Chien I, Regan T, et al. Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. PLoS ONE. 2023;18(11):e0272685. [doi: 10.1371/journal.pone.0272685] [Medline: 38011176]
18. Hilbert K, Böhnlein J, Meinke C, et al. Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: a machine learning study in two large multi-site samples in anxiety disorders. Neuroimage. Jul 15, 2024;295(120639):120639. [doi: 10.1016/j.neuroimage.2024.120639] [Medline: 38796977]
19. Ohi K, Fujikane D, Takai K, et al. Clinical features and genetic mechanisms of anxiety, fear, and avoidance: a comprehensive review of five anxiety disorders. Mol Psychiatry. Oct 2025;30(10):4928-4936. [doi: 10.1038/s41380-025-03155-1] [Medline: 40830577]
20. Basterfield C, Newman MG. Development of a machine learning-based multivariable prediction model for the naturalistic course of generalized anxiety disorder. J Anxiety Disord. Mar 2025;110:102978. [doi: 10.1016/j.janxdis.2025.102978] [Medline: 39904097]
21. Cruz-Gonzalez P, He AWJ, Lam EP, et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med. Feb 6, 2025;55:e18. [doi: 10.1017/S0033291724003295] [Medline: 39911020]
22. 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]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. Jan 26, 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. 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]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. Feb 18, 2014;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res. Dec 2023;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
27. Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology. Apr 2014;39(5):1254-1261. [doi: 10.1038/npp.2013.328] [Medline: 24270731]
28. Bertie LA, Quiroz JC, Berkovsky S, et al. Predicting remission following CBT for childhood anxiety disorders: a machine learning approach. Psychol Med. Dec 17, 2024;17:1-11. [doi: 10.1017/S0033291724002654] [Medline: 39686883]
29. Frick A, Engman J, Alaie I, et al. Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. J Affect Disord. Jan 15, 2020;261:230-237. [doi: 10.1016/j.jad.2019.10.027] [Medline: 31655378]
30. Hahn T, Kircher T, Straube B, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry. Jan 2015;72(1):68-74. [doi: 10.1001/jamapsychiatry.2014.1741] [Medline: 25409415]
31. Hentati Isacsson N, Ben Abdesslem F, Forsell E, Boman M, Kaldo V. Methodological choices and clinical usefulness for machine learning predictions of outcome in Internet-based cognitive behavioural therapy. Commun Med (Lond). Oct 10, 2024;4(1):196. [doi: 10.1038/s43856-024-00626-4] [Medline: 39384934]

32. Månsson KNT, Frick A, Boraxbekk CJ, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl Psychiatry. Mar 17, 2015;5(3):e530. [doi: 10.1038/tp.2015.22] [Medline: 25781229]
33. Sundermann B, Bode J, Lueken U, et al. Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia. Front Psychiatry. 2017;8:99. [doi: 10.3389/fpsyt.2017.00099] [Medline: 28649205]
34. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A, et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry. May 2015;21(5):680-685. [doi: 10.1038/mp.2015.109] [Medline: 26260493]
35. Vieira S, Liang X, Guiomar R, Mechelli A. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. Clin Psychol Rev. Nov 2022;97:102193. [doi: 10.1016/j.cpr.2022.102193] [Medline: 35995023]
36. Jankowsky K, Krakau L, Schroeders U, Zwerenz R, Beutel ME. Predicting treatment response using machine learning: a registered report. Br J Clin Psychol. Jun 2024;63(2):137-155. [doi: 10.1111/bjc.12452] [Medline: 38111213]
37. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Digit Health. 2021;7:20552076211060659. [doi: 10.1177/20552076211060659] [Medline: 34868624]
38. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry. Jun 2021;20(2):154-170. [doi: 10.1002/wps.20882] [Medline: 34002503]
39. Piccolo SR, Mecham A, Golightly NP, Johnson JL, Miller DB. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. PLoS Comput Biol. Mar 2022;18(3):e1009926. [doi: 10.1371/journal.pcbi.1009926] [Medline: 35275931]
40. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. Sep 25, 2024;26:e59505. [doi: 10.2196/59505] [Medline: 39321458]
41. Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P. Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. BMJ. Feb 14, 2012;344:e813. [doi: 10.1136/bmj.e813] [Medline: 22334559]
42. Fomenko A, Dümmler D, Aktürk Z, et al. Hospital Anxiety and Depression Scale Anxiety subscale (HADS-A) for detecting anxiety disorders in adults. Cochrane Database Syst Rev. Jul 2, 2025;7(7):CD015456. [doi: 10.1002/14651858.CD015456] [Medline: 40600405]
43. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry. Jun 2023;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
44. Abrol A, Fu Z, Salman M, et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun. Jan 13, 2021;12(1):353. [doi: 10.1038/s41467-020-20655-6] [Medline: 33441557]
45. Kartoun U, Khurshid S, Kwon BC, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. Sci Rep. Jul 22, 2022;12(1):12542. [doi: 10.1038/s41598-022-16615-3] [Medline: 35869152]
46. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Stat Med. Jul 10, 2021;40(15):3533-3559. [doi: 10.1002/sim.8981] [Medline: 33948970]
47. Benkarim O, Paquola C, Park BY, et al. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol. Apr 2022;20(4):e3001627. [doi: 10.1371/journal.pbio.3001627] [Medline: 35486643]
48. Li Y, Yu C, Zhao Y, Yao W, Aseltine RH, Chen K. Pursuing sources of heterogeneity in modeling clustered population. Biometrics. Jun 2022;78(2):716-729. [doi: 10.1111/biom.13434] [Medline: 33527347]
49. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. Sep 22, 2023;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
50. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. Computer Methods and Programs in Biomedicine Update. 2024;5:100148. [doi: 10.1016/j.cmpbup.2024.100148]
51. Hammelrath L, Hilbert K, Heinrich M, Zagorscak P, Knaevelsrud C. Select or adjust? How information from early treatment stages boosts the prediction of non-response in internet-based depression treatment. Psychol Med. Jun 2024;54(8):1641-1650. [doi: 10.1017/S0033291723003537] [Medline: 38087867]
52. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. Mar 15, 2018;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]

53. Tendolkar VD, Chouhan DS, Roy D, et al. Advances in evidence-based interventions for depression and anxiety: implications for mental health nursing practice. Int J Environ Sci. 2025;11(23s):826-834. URL: https://theaspd.com/index.php/ijes/article/view/8382/6055 [Accessed 2026-03-10] [doi: 10.64252/mev1b885]
54. As'ad M, Faran N, Joharji H. AI-Supported Shared Decision-Making (AI-SDM): conceptual framework. JMIR AI. Aug 7, 2025;4:e75866. [doi: 10.2196/75866] [Medline: 40773762]
55. Manole A, Cârciumaru R, Brînzaș R, Manole F. An exploratory investigation of chatbot applications in anxiety management: a focus on personalized interventions. Information. 2024;16(1):11. [doi: 10.3390/info16010011]
56. Das KP, Gavade P. A review on the efficacy of artificial intelligence for managing anxiety disorders. Front Artif Intell. 2024;7:1435895. [doi: 10.3389/frai.2024.1435895] [Medline: 39479229]
57. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology. Jun 2025;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CBT: cognitive behavioral therapy
CGI-I: Clinical Global Impressions–Improvement
DL: deep learning
FN: false negative
FP: false positive
GAD: generalized anxiety disorder
GAD-7: Generalized Anxiety Disorder-7
HAM-A: Hamilton Anxiety Rating Scale
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
LSAS: Liebowitz Social Anxiety Scale
ML: machine learning
OASIS: Overall Anxiety Severity and Impairment Scale
OCD: obsessive-compulsive disorder
PD: panic disorder
PHQ-9: Patient Health Questionnaire-9
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence
PTSD: posttraumatic stress disorder
RF: random forest
RNN: recurrent neural network
SAD: social anxiety disorder
SPQ: Spider Phobia Questionnaire
SVM: support vector machine
TN: true negative
TP: true positive

Edited by Stefano Brini; peer-reviewed by Charlotte Ahmadu, Dhrubajyoti Ghosh, Sadhasivam Mohanadas, Sandesh Gubbi Virupaksha, Yonggang Zhang; submitted 17.Oct.2025; final revised version received 12.Feb.2026; accepted 24.Feb.2026; published 18.Mar.2026

Please cite as:
Liu J, Wang J, Wu Z, Adam Assim MISB
Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e86079
URL: https://www.jmir.org/2026/1/e86079
doi: 10.2196/86079