Review

Machine Learning for Comparative Antidepressant Selection in Major Depressive Disorder: Systematic Review

Fiona He[1], BA; Steven Huang[2,3]; Richard Wang[4]; Aland Chang[2]; Jennifer L Phillips[5], PhD; Christopher Sun[3,6], PhD

[1]Digital Transformation and Innovation, Faculty of Engineering, University of Ottawa, Ottawa, ON, Canada
[2]Faculty of Science, University of British Columbia, Vancouver, BC, Canada
[3]Ottawa Heart Institute, Ottawa, ON, Canada
[4]Department of Physiology and Pharmacology, Western University, London, ON, Canada
[5]University of Ottawa Institute of Mental Health Research at The Royal, Ottawa, ON, Canada
[6]Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:

Christopher Sun, PhD
Telfer School of Management
University of Ottawa
55 Laurier Ave E
Ottawa, ON K1N 9B9
Canada
Phone: 1 (613) 562-5800 ext 4570
Email: sun@telfer.uottawa.ca

Abstract

Background: Major depressive disorder (MDD) affects approximately 1 in 6 adults during their lifetime, yet antidepressant selection relies predominantly on trial-and-error, with response rates of only 42% to 53%. While machine learning (ML) models have shown promise in predicting treatment outcomes, most focus on single treatments rather than comparative selection across therapeutic alternatives, limiting their clinical utility for the medication choice decisions that clinicians face in practice.

Objective: This systematic review evaluates ML approaches that examine 2 or more pharmacological interventions for predicting treatment outcomes in MDD, with a focus on their capacity to facilitate comparative treatment selection between medications or medication classes for individual patients.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we searched PubMed, Scopus, and Web of Science for studies published from 2015 to 2025. We included studies involving adults with MDD that used ML models to predict treatment outcomes across 2 or more pharmacological treatments and reported medication-specific prediction outcomes. Risk of bias was assessed using PROBAST-AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence). We conducted a narrative synthesis organized by modeling strategies, data integration approaches, validation methodologies, and performance patterns.

Results: From 5370 initial records, 19 studies met the inclusion criteria, with dataset sample sizes ranging from 49 to 77,226 participants. Studies employed 3 distinct modeling strategies: drug-specific supervised models trained independently for each medication, subtype- or trajectory-based approaches using clustering methods to identify differential response patterns, and a unified differential prediction framework generating calibrated cross-treatment predictions. Performance varied substantially, with area under the curve values ranging from 0.59 to 0.95 and classification accuracies between 62% and 95.4%, though high performance was concentrated in studies with small samples, high-dimensional neurobiological features, and internal-only validation. Only 7 studies conducted external validation, which generally yielded more conservative performance estimates. Feature informativeness was more consistently associated with performance variation than algorithm complexity. Most studies did not formally distinguish between prognostic features predicting general outcomes and predictive features identifying differential medication responses, and none applied formal explainability techniques.

Conclusions: ML for comparative antidepressant selection remains in an early stage of development. Only 1 study implemented a unified framework directly supporting patient-level treatment ranking. Key barriers to clinical translation include insufficient distinction between prognostic and predictive markers, limited cross-trial validation, near-absent calibration

reporting, and absent explainability. Future research should prioritize unified comparative frameworks with calibrated predictions, rigorous external validation on diverse cohorts, explicit modeling of heterogeneous treatment effects, and integration of explainability into model development.

JMIR Ment Health 2026;13:e89352; doi: 10.2196/89352

Keywords: machine learning; artificial intelligence; antidepressant; major depressive disorder; treatment selection; treatment outcome; precision psychiatry; personalized medicine

Introduction

Major depressive disorder (MDD) represents one of the most prevalent mental health conditions globally, impacting 1 in 6 adults during their lifetime [1]. The disorder is characterized by persistent depressed mood, anhedonia, and a variety of physical and cognitive symptoms that significantly impair psychosocial functioning and diminish health-related quality of life [2,3]. When the condition becomes severe, patients may exhibit suicidal or self-harming tendencies, increasing the risk of adverse outcomes and health care system burden [4].

Despite the availability of numerous pharmacological interventions, selecting effective antidepressants for individual patients remains challenging due to substantial heterogeneity in symptom presentation, disease progression, and individual treatment response patterns [5,6]. Current antidepressant selection predominantly follows a trial-and-error paradigm, with clinicians initiating treatment based on clinical guidelines and subsequently adjusting medication plans according to observed patient responses [6]. This approach yields modest response rates of approximately 42% to 53%, meaning that nearly half of patients do not achieve adequate symptom improvement, often resulting in prolonged time to remission, poor clinical outcomes, and elevated treatment discontinuation rates [6,7]. These therapeutic inefficiencies also impose significant societal economic costs through increased health care utilization and loss of work productivity [8,9].

The emergence of precision medicine in psychiatry has catalyzed growing interest in leveraging machine learning (ML) methodologies to optimize pharmacotherapy selection for MDD. ML, a subfield of artificial intelligence, encompasses computational methods that enable systems to learn from data and improve performance through experience. ML algorithms can process complex, high-dimensional, and multimodal health care datasets, including electronic health records, genetic markers, neuroimaging data, and clinical assessments, while identifying subtle patterns predictive of treatment outcomes that may not be detected through conventional analytical approaches [10-12]. Recent studies have demonstrated promising results in applying ML to predict treatment outcomes [6,10,11,13-17]. Compared to traditional statistical methods, ML technologies offer enhanced scalability and adaptability, and demonstrate potential for more accurate prediction of individual patient responses to specific therapeutic interventions [18].

The fundamental clinical challenge extends beyond predicting treatment response. Clinical practice inherently involves comparative decision-making. Clinicians must choose which medication to initiate, when to switch to an alternative agent, or whether to augment with additional pharmacotherapy, with each option presenting different mechanisms of action, side effect profiles, and patient-specific considerations [19,20]. A substantial proportion of patients who do not remit on an initial antidepressant achieve remission after switching, suggesting that early nonresponse may reflect suboptimal treatment matching rather than true treatment resistance [21,22]. Prediction models that estimate response to a single medication in isolation cannot directly support this comparative decision-making [23]. Effective treatment selection requires approaches that compare expected outcomes across therapeutic alternatives for individual patients, identifying which specific medication is most likely to provide superior benefit, the essential component for personalized medication selection. In addition, limited model explainability remains a critical barrier to clinical translation. Many ML algorithms function as “black boxes,” providing predictions without transparent reasoning about which patient characteristics drive recommendations [24,25]. For comparative treatment selection, this limitation is particularly concerning because clinicians must understand not only which medication is recommended but also why one option may be preferred over alternatives for a specific patient, enabling them to integrate model outputs with clinical judgment and patient preferences.

In this review, our primary goal is to evaluate the current state of ML approaches that facilitate comparative treatment selection between different medications or medication classes for personalized pharmacotherapy management in MDD. We examine key characteristics of existing models, including their comparative capabilities, data integration approaches, validation methods, and clinical applicability. By systematically evaluating these aspects, this review seeks to identify current limitations and provide insights for developing ML models that can meaningfully guide clinical decision-making in precision psychiatry.

Methods

Search Strategy

This systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26]. We performed a search across PubMed, Scopus, and Web of Science databases for all related studies published from 2015 to 2025. The search strategy combined three categories of terms: (1) ML and related approaches, (2) treatment response and outcome

prediction, and (3) depression and antidepressant pharmacotherapy, including individual drug classes and specific medication names. The complete search strings for each database are provided in Multimedia Appendix 1.

Eligibility Criteria

Four independent reviewers screened titles and abstracts, followed by a full-text review of potentially eligible studies. We included original studies involving adult patients diagnosed with MDD that used ML models for predicting treatment outcomes or supporting personalized pharmacotherapy management. Studies were required to examine 2 or more pharmacological treatments or different dosage levels of the same antidepressant, and to report medication-specific prediction outcomes that enable comparison between different antidepressants for individual patients. We excluded studies that examined single treatments without comparison, those focusing only on side effects, nonpharmacological interventions, secondary depression, and those published in languages other than English.

For this review, we defined "comparative treatment selection capability" as a model's ability to inform individual-level medication choices between different antidepressants. Studies were categorized based on their approaches to handling multiple treatments.

Data Extraction and Synthesis

Data extraction was conducted independently by reviewers using a standardized extraction form. Discrepancies were reviewed by a third senior reviewer, who made the final decision after discussion with the team. We extracted study information related to study design and ML methodology. Table 1 documents basic study characteristics, including population, study outcome, outcome measurement, and prediction time horizon. Tables 2 and 3 focused on the methodological details of ML, capturing algorithm types, feature selection, validation strategies, and performance metrics. Risk of bias was assessed using the PROBAST-AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence). Two reviewers independently rated each study across 4 domains (participants, predictors, outcome, and analysis), with disagreements resolved through discussion [27,28].

Table 1. Experimental design characteristics of the included studies (n=19).

Author	Population (n)	Outcome category	Outcome measure	Prediction horizon (wk)	Intervention
Iniesta et al [29] (2016)	GENDEP[a] study (793)	Treatment response	HAM-D[b], MADRS[c], and BDI[d] remission	12	SSRI[e] (escitalopram), TCA[f] (nortriptyline), combination
Chekroud et al [30] (2017)	STAR*D[g] + CO-MED[h] (4706) + duloxetine (2515); Total (7221)	Medication recommendation	HAM-D for 8 wk and QIDS-SR[i] for 12 wk (symptom clusters)	8-12	SSRI (escitalopram) + Placebo, SSRI (escitalopram) + NDRI[j] (bupropion), SNRI[k] (venlafaxine) + NaSSA[l] (mirtazapine), SNRI (duloxetine) + SSRIs + Placebo
Crane et al [31] (2017)	University of Michigan (49)	Treatment response	HAM-D (%[m] change; Hamilton Depression Rating Scale pretreatment to posttreatment)	10	SSRI (escitalopram), SNRI (duloxetine)
Iniesta et al [32] (2018)	GENDEP study (n=430)	Treatment remission	17-item HAM-D[n]	4-12	SSRI (escitalopram) and SNRI (nortriptyline)
Kautzky et al [33] (2021)	GGRND[o] study (1070)	Treatment response and remission	HAM-D (treatment response=baseline score >50% and remission <7)	8	SSRI, TCA, antipsychotics, lithium augmentation
Hughes et al [34] (2020)	Site A (51,048), site B (26,176), total (77,226)	Treatment stability	Treatment stability (≥2 prescriptions, ≥30 d apart, ≥90 d duration, MPR ≥80%)	12	SSRI (citalopram, sertraline, fluoxetine, escitalopram, paroxetine), SNRI (venlafaxine, duloxetine), NDRI (bupropion), TCA (nortriptyline, amitriptyline), NaSSA (mirtazapine)
Taliaz et al [6] (2021)	STAR*D (4041)	Treatment response	HAM-D and QIDS[p] (≥50% reduction)	12-14	SSRI (citalopram, sertraline), SNRI (venlafaxine)
Athreya et al [35] (2021)	PGRN-AMPS[q] + ISPC[r] (947)	Treatment response	HAM-D (treatment response=baseline score >50% and remission total score ≤7)	8	SSRI (citalopram, escitalopram)
Bi et al [36] (2021)	West China Hospital of Sichuan University (610)	Treatment response	HAM-D (treatment response=baseline score >50% and remission total score <8)	6	SSRI (fluoxetine, paroxetine, citalopram, sertraline), SNRI (duloxetine, venlafaxine), TCA (amitriptyline, doxepin, imipramine), NaSSA (mirtazapine)
Nguyen et al [37] (2022)	EMBARC[s] study (222)	Treatment response	HAM-D (treatment response ≥50% from pre-treatment and remission ≤7)	8-16	SSRI (sertraline), NDRI (bupropion), placebo

Author	Population (n)	Outcome category	Outcome measure	Prediction horizon (wk)	Intervention
Wang et al [38] (2022)	Multihospital study (430)	Early improvement	20% reduction in 17-item HAMD score	2	SSRIs, SNRIs, combinations, rTMS[t] + antidepressant, and ECT[u] + antidepressant
Chen et al [39] (2023)	310	Treatment response	QIDS-SR-16 and HAM-D28	10-12	SSRI (escitalopram), NDRI (bupropion) or combination (escitalopram + bupropion), SNRI (duloxetine), or placebo
Turner et al [40] (2023)	P64808 and UMCU[v] (735)	Medication recommendation	Acceptability (prescription duration ≥5 wk) and efficacy (NLP[w]-extracted recovery themes)	14 and 23	SSRI (sertraline, citalopram, escitalopram, fluoxetine, paroxetine, fluvoxamine), SNRI (trazodone, duloxetine, venlafaxine), TCA (amitriptyline, clomipramine, imipramine, doxepin, maprotiline, dosulepine), MAOI[x] (tranylcypromine, moclobemide, phenelzine), other (bupropion, vortioxetine, agomelatine, hypericum herb, mirtazapine, mianserine)
Fu et al [16] (2024)	COORDINATE-MDD[y] (1384)	Treatment response	17-item HAM-D	6, 8, or 12	SSRI (sertraline, citalopram, escitalopram), placebo
Curtiss et al [14] (2024)	STAR*D Stage 2 (1439)	Remission	HAM-D (remission = score < 7)	12	SSRI (sertraline, citalopram), SNRI (venlafaxine), NDRI (bupropion), cognitive psychotherapy
Ravan et al [41] (2024)	EMBARC study (224)	Treatment response	17-item HAM-D (treatment response ≥50%)	8	SSRI (sertraline), NDRI (bupropion), placebo
Benrimoh et al [42] (2025)	Pooled antidepressant clinical trials (9042)	Treatment remission	MADRS<11, QIDS-SR-16<6, or HAMD<8	6-14	10 pharmacological treatments
Carr et al [43] (2025)	GENDEP (714)	Treatment response	HAM-D (remission=score ≤ 7)	12	SSRI (escitalopram), TCA (nortriptyline)
Zhukovsky et al [44] (2025)	EMBARC and Canadian Biomarker Integration Network in Depression-1 (363)	Treatment response	≥50% reduction in depression severity (HDRS or converted MADRS)	8, 16	SSRIs (sertraline and escitalopram)

[a]GENDEP: Genome-Based Therapeutic Drugs for Depression.
[b]HAM-D: Hamilton Depression Rating Scale.
[c]MADRS: Montgomery-Åsberg Depression Rating Scale.
[d]BDI: Beck Depression Inventory.
[e]SSRI: selective serotonin reuptake inhibitor.
[f]TCA: tricyclic antidepressant.
[g]STAR*D: Sequenced Treatment Alternatives to Relieve Depression.
[h]CO-MED: combining medications to enhance depression outcomes.
[i]QIDS-SR: Quick Inventory of Depressive Symptomatology-Self Report version.
[j]NDRI: norepinephrine-dopamine reuptake inhibitor.
[k]SNRI: serotonin-norepinephrine reuptake inhibitor.
[l]NaSSA: noradrenergic and specific serotonergic antidepressant.
[m]HAM-D%: percent change in HAM-D score.
[n]HAM-D17: 17-item Hamilton Depression Rating Scale.
[o]GRND: Genomic-based Response to Depression dataset.
[p]QIDS: Quick Inventory of Depressive Symptomatology.
[q]PGRN-AMPS: Pharmacogenomics Research Network-Antidepressant Medication Pharmacogenomics Study.
[r]ISPC: International SSRI Pharmacogenomics Consortium.
[s]EMBARC: Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care.
[t]rTMS: repetitive transcranial magnetic stimulation.
[u]ECT: electroconvulsive therapy.
[v]UMCU: University Medical Center Utrecht.
[w]NLP: natural language processing.
[x]MAOI: monoamine oxidase inhibitor.
[y]COORDINATE-MDD: Coordinated Outcomes in Depression for Individualized and Translational Evaluation in Major Depressive Disorder.

Table 2. Feature modality and validation strategy across included studies (n=19).

Feature modality	Reference	Internal validation	External validation
Clinical-only	• Iniesta et al [29] (2016) • Chekroud et al [30] (2017) • Kautzky et al [33] (2021) • Hughes et al [34] (2020) • Athreya et al [35] (2021) • Chen et al [39] (2023) • Turner et al [40] (2023) • Curtiss et al [14] (2024) • Benrimoh et al [42] (2025)	• Iniesta et al [29] (2016) • Chekroud et al [30] (2017) • Kautzky et al [33] (2021) • Turner et al [40] (2023) • Curtiss et al [14] (2024) • Benrimoh et al [42] (2025)	• Chekroud et al [30] (2017) • Hughes et al [34] (2020) • Athreya et al [35] (2021) • Chen et al [39] (2023)
Neurobiological-only	• Fu et al [16] (2024) • Ravan et al [41] (2024)	• Fu et al [16] (2024) • Ravan et al [41] (2024)	• No
Multimodal	• Crane et al [31] (2017) • Iniesta et al [32] (2018) • Taliaz et al [6] (2021) • Bi et al [36] (2021) • Nguyen et al [37] (2022) • Wang et al [38] (2022) • Carr et al [43] (2025) • Zhukovsky et al [44] (2025)	• Iniesta et al [32] (2018) • Crane et al [31] (2017) • Nguyen et al [37] (2022) • Carr et al [43] (2025)	• Taliaz et al [6] (2021) • Wang et al [38] (2022) • Zhukovsky et al [44] (2025)

Table 3. Machine learning approaches and predictive performance across included studies (n=19).

Category and reference	Machine learning techniques	Model type	Performance metrics
Drug-specific supervised models			
Iniesta et al [29] (2016)	ENRR[a]	Classification, regression	Explained 5%-10% variance in symptom improvement; Remission AUC[b]=0.75 for escitalopram and AUC=0.72 for nortriptyline.
Crane et al [31] (2017)	ICA[c] + multiple regression/RF[d]	Regression	Remission prediction accuracy improved from 74% to 90% after adding fMRI[e] features.
Iniesta et al [32] (2018)	ENRR	Classification	Remission prediction AUC=0.77 for escitalopram and nortriptyline
Taliaz et al [6] (2021)	SVM[f] (linear kernel), XGBoost[g], RF	Classification	Average balanced accuracy of 70.1% across medications in the final test set.
Bi et al [36] (2021)	RF (feature selection) + multiple generalized regression	Classification	Prediction model AUC=77% for SSRI[h] and 75% for SNRI[i].
Nguyen et al [37] (2022)	Deep learning (feed-forward neural networks) with data augmentation	Classification, regression	R^2 of 48% for Sertraline and 34% for Bupropion in predicting symptom change.
Ravan et al [41] (2024)	CNN[j] + symbolic transfer entropy (STE[k]) + ReLORETA[l]	Classification	Overall prediction accuracy >85% (Sertraline: 91%; Placebo: 95.4%; Bupropion: 86.8%).
Curtiss et al [14] (2024)	Super learner ensemble (RF, Elastic Net, NN, XGBoost, etc)	Classification	Prediction AUC=0.51-0.82 (highest for cognitive therapy at 0.82; Bupropion at 0.70).
Zhukovsky et al [44] (2025)	Elastic net logistic regression and multivariate partial least squares regression (PLS-R)	Classification, regression	Escitalopram AUC=0.66 (balanced accuracy=0.64) Sertraline AUC=0.70 (balanced accuracy=0.71)
Subtype or trajectory-based models			
Chekroud et al [30] (2017)	Hierarchical clustering + linear mixed effects regression + gradient boosting	Regression	Explained variance for symptom clusters: sleep (19.6%), core emotional (14.5%), atypical (15.1%).
Hughes et al [34] (2020)	Prediction-constrained topic modeling + extremely randomized trees/logistic regression	Classification	General stability prediction AUC=0.627-0.661; drug-specific models performed similarly.
Kautzky et al [33 (2021)	Hierarchical symptom clustering + random forest	Classification	Max prediction accuracy of 0.85 (Cluster IV); SSRI-stratified group accuracy=0.82.
Athreya et al [35] (2021)	GMM[m] + HMM[n] + PGM[o]	Classification	Average 8-wk outcome prediction accuracy: SSRI group=77%; Other drugs=72%.
Wang et al [38] (2022)	Hierarchical clustering combined with canonical correlation analysis (CCA) to	Classification	Achieved an overall cross-dataset accuracy of 72.83%, SSRI accuracy=83.11%, SNRIs accuracy=66.67%

Category and reference	Machine learning techniques	Model type	Performance metrics
	link functional connectivity to clinical symptoms.		
Turner et al [40] (2023)	NLP[p] + BN[q]	Network analysis	Discovered 28 dependencies between features and outcomes; continuation rates were 66% and 89%
Chen et al [39] (2023)	Hierarchical clustering + penalized logistic regression	Regression	Sleep cluster: R^2=45%, RMSE=81; Atypical cluster: R^2=41%[r], RMSE[s]=85; Core emotional: R^2=42%, RMSE=92
Fu et al [16] (2024)	Nonlinear semisupervised clustering (HYDRA[t]) + SVM	Classification	Cluster stability ARI[u]=0.61; significant dimension-by-treatment interaction.
Carr et al [43] (2025)	GC[v] + TDA[w] + elastic net logistic regression	Classification	Week 4 prediction AUC: escitalopram (0.807), nortriptyline (0.777), combined sample (0.794).
Unified differential prediction model			
Benrimoh et al [42] (2025)	Deep learning	Classification	AUC of 0.65 on the held-out test set. Simulation analysis estimated an increase in population remission rates from 43.63% to 53.98%

[a]ENRR: elastic net regularized regression.
[b]AUC: area under the curve.
[c]ICA: independent component analysis.
[d]RF: random forest.
[e]fMRI: functional magnetic resonance imaging.
[f]SVM: support vector machine.
[g]XGBoost: eXtreme Gradient Boosting.
[h]SSRI: Selective Serotonin Reuptake Inhibitor.
[i]SNRI: Serotonin-Norepinephrine Reuptake Inhibitor.
[j]CNN: convolutional neural network.
[k]STE: symbolic transfer entropy.
[l]ReLORETA: robust exact low-resolution electromagnetic tomography.
[m]GMM: Gaussian mixture models.
[n]HMM: hidden Markov model.
[o]PGM: probabilistic graphical models.
[p]NLP: natural language processing.
[q]BN: Bayesian network.
[r]R^2: coefficient of determination.
[s]RMSE: root mean square error.
[t]HYDRA: heterogeneity through discriminative analysis
[u]ARI: adjusted rand index (a measure of cluster similarity/stability).
[v]GC: growth curves.
[w]TDA: topological data analysis.

We conducted a narrative synthesis, organizing findings by: (1) modeling strategies for handling multiple treatments, (2) data integration approaches, (3) validation methodologies, and (4) performance across different antidepressant classes.

Risk of Bias Assessment

Risk of bias was assessed using an adaptation of PROBAST incorporating AI-specific signaling questions for ML prediction models [27,28]. The assessment evaluated 4 domains: participants, predictors, outcome, and analysis, with additional considerations for sample size adequacy, high-dimensional predictor handling, class imbalance correction, and overfitting prevention. Separate assessments were conducted for model development and model validation phases. Two reviewers independently rated each study, with disagreements resolved through discussion. All studies were retained regardless of risk of bias classification, given the limited eligible literature. Complete assessment results are reported in Multimedia Appendix 2.

Results

Description of the Included Studies

Our initial search yielded a total of 5370 studies. After removing duplicates and applying inclusion criteria, 19 studies remained (Figure 1).

The 19 included studies demonstrated significant variation in dataset sample sizes for ML model development, ranging from 49 to 77,226 participants. Multiple studies utilized well-established clinical datasets, with considerable overlap across investigations. The Sequenced Treatment Alternatives to Relieve Depression, Genome-Based Therapeutic Drugs for Depression, and Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care datasets were the most frequently used, with each appearing in three studies [6,14,29,30,32,37,41,43,44]. This indicates significant reuse of established research cohorts. Geographically, the majority of studies originated from Western countries, particularly the United States and European institutions, while 2 investigations were conducted in China [36,38]. This distribution

reflects the current concentration of ML research infrastructure in established academic medical centers (Figure 2).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion and exclusion criteria. ML: machine learning.

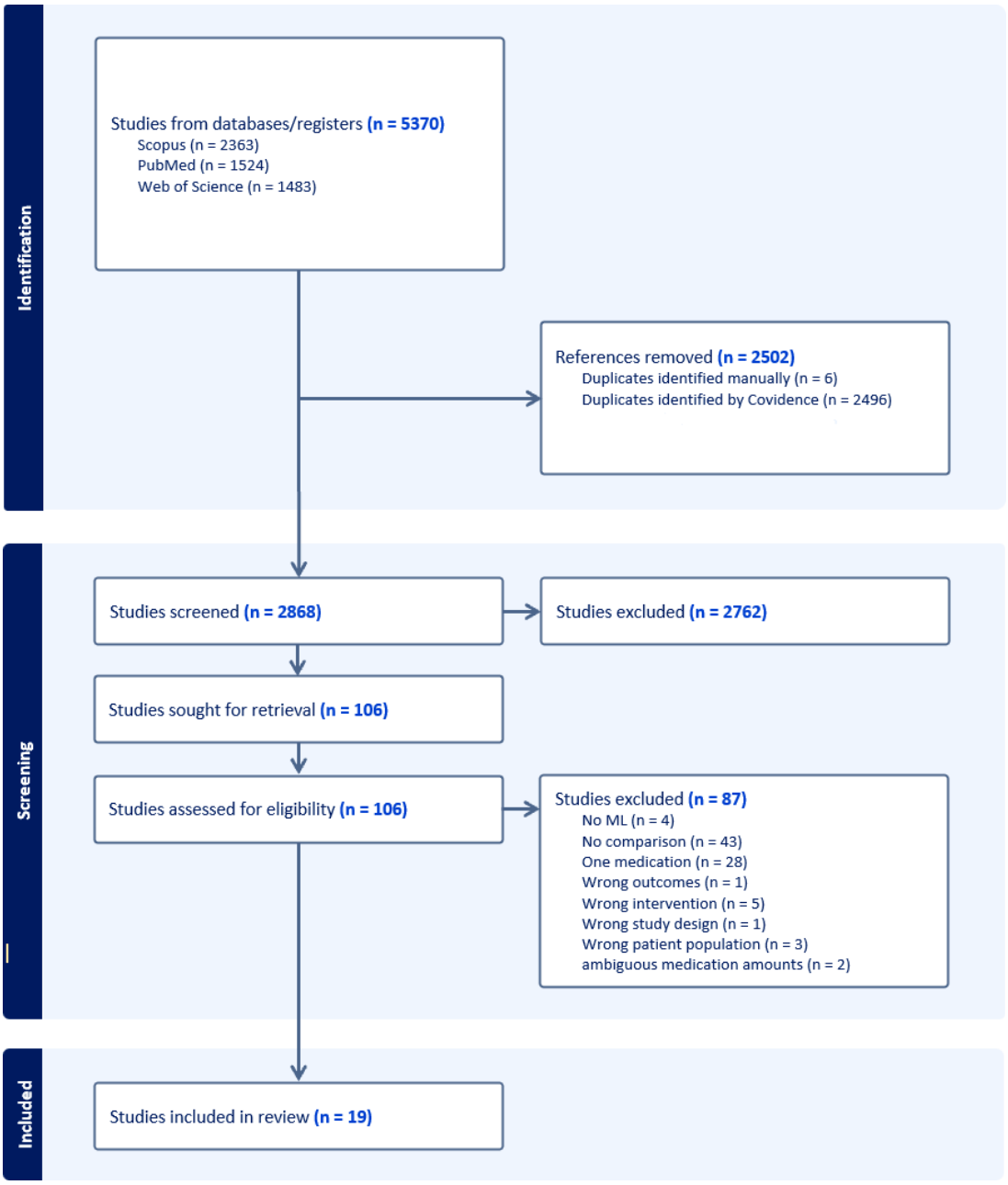
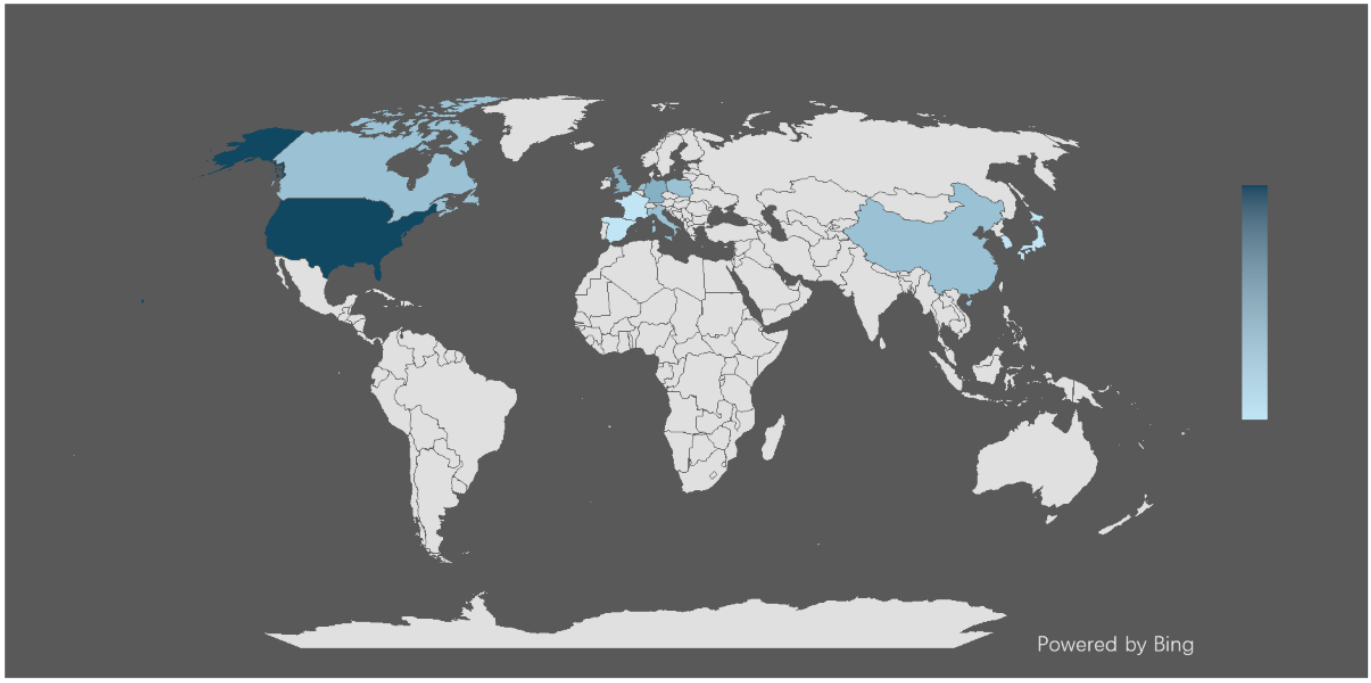


Figure 2. Geographic distribution of datasets used in the included studies (n=19). Figure created in Microsoft Word using the built-in map chart function [45], which is published under limited license per the Microsoft Bing Maps Terms of Use [46].



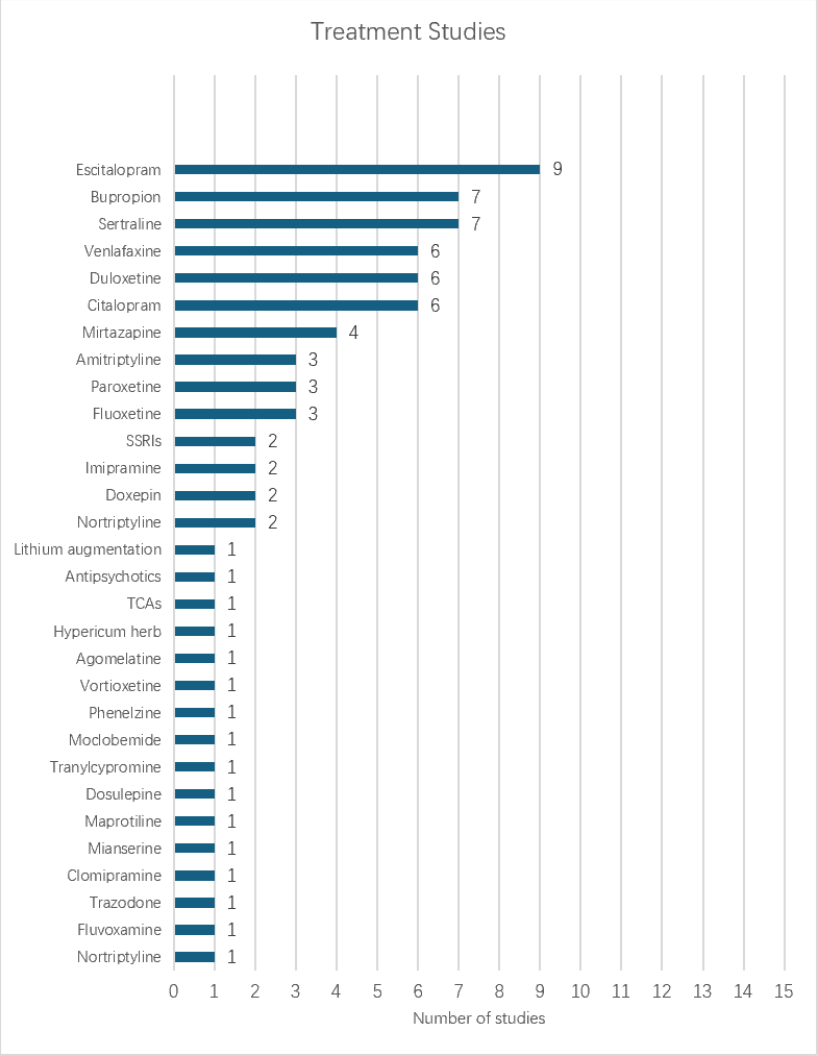
Data Input, Feature Modality, and Validation Strategy

Most studies relied on structured clinical and demographic data, including symptom ratings, treatment history, and comorbidities [14,29,30,33-35,39,40,42]. Neurobiological features were examined in 2 studies, using pretreatment electroencephalogram connectivity and structural magnetic resonance imaging, respectively [16,41]. The remaining 8 studies pursued multimodal integration, combining neurobiological or genetic data with clinical variables [6,31,32,36-38,43,44]. Turner et al [40] introduced natural language processing methods to derive structured predictors from unstructured clinical notes, highlighting an alternative pathway for utilizing real-world data.

Internal validation was the predominant approach across studies. Only 7 studies conducted external validation on independent datasets or cross-site samples, though the scope varied considerably [6,30,34,35,38,39,44]. Five studies validated their full modeling framework on independent cohorts, while Zhukovsky et al [44] conducted cross-trial validation between 2 randomized controlled trials (RCTs). Taliaz et al [6] conducted external validation only on a single medication model. Among these externally validated studies, 4 relied on clinical-only features, suggesting that external validation has been predominantly pursued in studies using routinely available data. This pattern, alongside the tendency for neurobiological studies to use smaller cohorts with internal validation only, precludes direct performance comparisons across feature modality categories.

Antidepressant coverage was also uneven (Figure 3). Escitalopram and bupropion were most frequently modeled, followed by sertraline, venlafaxine, duloxetine, and citalopram. A small number of studies additionally examined treatment classes (eg, Selective Serotonin Reuptake Inhibitors [SSRIs], tricyclic antidepressants) or augmentation strategies (eg, lithium).

Figure 3. Distribution of antidepressant treatments modeled across included studies. SSRI: Selective Serotonin Reuptake Inhibitor; TCA: tricyclic antidepressant.



Explainability and Interpretability

Most of the included studies did not apply formal explainable artificial intelligence techniques, except for Benrimoh et al [42] and Nguyen et al [37]. Benrimoh et al [42] used gradient-based saliency maps using the GuidedBackprop algorithm to generate patient-level interpretability reports, providing clinicians with the top 5 features driving each individual treatment prediction. Nguyen et al [37] used the partial derivative method to rank feature importance learned by their deep learning models, reporting the 20 most important predictive features for each treatment arm. Explainable artificial intelligence techniques are designed to provide transparent and interpretable explanations of ML model predictions, helping users understand how models arrive at specific decisions [47]. Beyond these 2 studies, model interpretability was either implicitly embedded within model architecture or addressed through basic post hoc statistical descriptions. Several studies employed inherently interpretable models, including logistic regression, Bayesian networks, or penalized linear models with coefficient inspection [29,32,39,40,43,44]. In many cases, these models were selected not explicitly for interpretability, but because they enabled effective feature selection for predicting antidepressant treatment response. Similarly, symptom-based clustering was employed to define patient subgroups with distinct response patterns, providing an indirect rationale for treatment matching [16,30,38,39].

Outcome Measures

The included studies demonstrated considerable variation in outcome measurement approaches. Depression severity was assessed using standardized rating scales in 17 of 19 (89.5%) studies. The Hamilton Depression Rating Scale was the most utilized measure across studies in various forms, including the 17-item and 28-item versions. Within this group, 3 studies combined the Hamilton Depression Rating Scale with the Quick Inventory of Depressive Symptomatology Self-Report, and 1 study additionally incorporated both the Montgomery-Åsberg Depression Rating Scale and the Beck Depression Inventory. The remaining 2 studies adopted alternative outcome measures. Hughes et al [34] evaluated treatment stability, and Turner et al [40] assessed treatment acceptability and efficacy.

Despite this general reliance on standardized scales, the definitions of treatment response varied substantially. Most studies defined response as a reduction of 50% or more in depressive symptoms on a selected scale, while some required a reduction of more than 50%. Wang et al [38] adopted a lower threshold of a reduction of 20% or more for defining early improvement at 2 weeks. Remission was typically characterized as achieving scores below predetermined scale-specific cutoff points, most commonly 7 or fewer.

The predictive time horizons, the periods over which outcomes are predicted, ranged from 2 weeks to 163 days. Most predicted time horizons were from 6 to 12 weeks. Only 1 study extended to 163 days for longer-term outcome assessment [40].

ML Techniques

A wide array of ML techniques has been employed across the included studies to predict antidepressant treatment outcomes (Table 3). Linear and logistic regression methods were frequently applied, including elastic net regression, which is valued for its feature selection and interpretability [29-32,34,39,43,44]. Tree-based methods, such as random forest and gradient boosting, were implemented in both stand-alone and ensemble configurations [30,33,34,36]. Support vector machines and deep learning models were less common. Deep learning was used in studies that incorporated high-dimensional neuroimaging or multimodal inputs, but was also applied to low-dimensional clinical and demographic data at scale [6,37,41,42]. Two studies used probabilistic graphical or network-based approaches to model relationships between patient features and treatment outcomes [35,40].

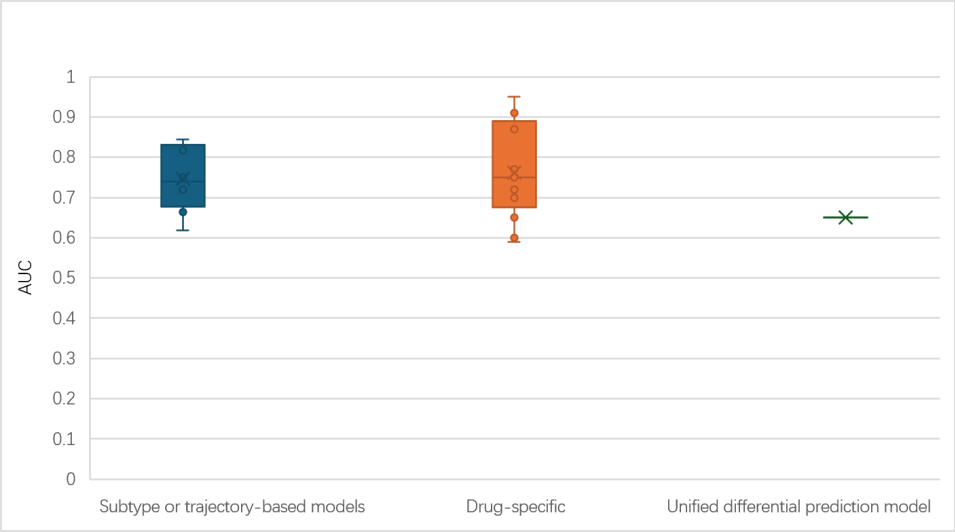
Across studies, the choice of algorithm broadly corresponded to the dimensionality of input data. Deep learning and multivariate regression were employed in studies with high-dimensional neuroimaging inputs [37,41,44]. Linear and penalized regression models were applied in studies using structured clinical variables, and tree-based or ensemble methods appeared in studies combining multiple data types [6,29,30,34,36,39].

In addition, the studies included varied in how they handled multiple pharmacological interventions. Several studies trained independent models for each antidepressant arm, applying the same algorithm and feature set separately, which enabled drug-specific predictions but did not support direct comparisons across treatments [6,14,29,31,32,36,37,41,44]. The second group of studies employed clustering or trajectory-based methods, grouping patients into subtypes or modeling treatment paths and then examining differential response patterns [30,33-35,38-40,43]. The third approach, represented by Benrimoh et al [42], developed a unified differential treatment prediction framework in which a single model simultaneously generated calibrated remission probabilities across 10 pharmacological treatments for each individual patient, enabling direct cross-treatment comparison. The training sample size also differed across approaches. Drug-specific models often had to rely on smaller effective cohorts per arm, and clustering approaches required sufficiently large subgroups to ensure stable results.

Performance

Reported performance metrics varied by medication arm, modeling strategy, and outcome type (Table 3). Studies demonstrated substantial heterogeneity in model performance, with area under the curve (AUC) values ranging from 0.59 to 0.95 for classification tasks and classification accuracies between 62% and 95.4% across conditions (Figure 4). However, such high performance levels were uncommon and concentrated in a small number of studies, while most reported more moderate predictive accuracy.

Figure 4. Comparison of area under the curve (AUC) across drug-specific supervised models, subtype- or trajectory-based models, and unified differential prediction model.



At the structural level, subtype- or trajectory-based models that did not distinguish between medications generally achieved lower but more stable performance compared with drug-specific approaches, which demonstrated greater

heterogeneity but included some of the highest reported values. For example, Hughes et al [34] reported AUC values of 0.627 to 0.661 for general stability prediction across external validation sites, whereas among drug-specific models, Ravan et al [41] achieved classification accuracies above 85% using electroencephalogram-based deep learning with internal validation only. This contrast illustrates the characteristic trade-off between generalizability and peak performance observed across the 2 approaches.

Incorporating early treatment response data as additional predictive features demonstrated consistent advantages across multiple studies. Carr et al [43] reported AUC values increasing from 0.703 at baseline to 0.794 at week 4 and 0.844 at week 6 in the combined sample. Similarly, Zhukovsky et al [44] found that replacing pretreatment depression severity with week 2 scores improved cross-trial AUC from 0.58 to 0.68 for the clinical-only model and up to 0.79 for the clinical plus neuroimaging model.

Four studies used regression frameworks to predict changes in continuous symptom severity [30,31,39,44]. Chen et al [39] applied penalized regression within symptom-based clusters, reporting R^2 values of up to 45% in the sleep cluster, with root mean square error values ranging from 81 to 92. Replicating the approach of Chekroud et al [30], these results contrasted with the lower values observed in the original models, where R^2 was below 0.20 [39]. Zhukovsky et al [44] used multivariate partial least squares regression incorporating functional connectivity features, reporting predicted-vs-observed correlations of 0.31 to 0.39 in cross-trial validation. Turner et al [40] used a Bayesian network approach to model probabilistic dependencies among patient features and treatment outcomes. Although not directly comparable to classification or regression models, their approach identified stable network structures and reported a 6% to 11% gain in treatment selection for specific subgroups. However, traditional performance metrics such as area under the receiver operating characteristic curve or root mean square error were not reported, which limits comparability.

Among the 3 modeling strategies, Benrimoh et al [42] reported the only unified differential prediction framework, achieving an AUC of 0.65 on the held-out test set across 10 pharmacological treatments. While this AUC was lower than many drug-specific models, the model's simulation analysis estimated an increase in population remission rates from 43.15% to 53.99% when treatment was selected based on model recommendations rather than random assignment.

Examining performance patterns across studies, several factors appeared systematically associated with variation in reported model performance. Feature complexity and data modality represented the most consistent source of performance differences. Studies incorporating neurobiological inputs, such as neuroimaging, electrophysiological signals, or genetic markers, generally reported higher predictive performance than those relying exclusively on clinical and demographic variables [31,41,43,44]. By comparison, studies using only structured clinical and demographic inputs reported more modest classification performance [14,29,30,33-35,40,42].

However, this feature-performance association covaried with sample size and validation approach in ways that complicate straightforward interpretation. Studies employing neurobiological features tended to use smaller research cohorts and relied predominantly on internal cross-validation. Conversely, the largest study in the review utilized structured clinical data and conducted external validation across independent clinical sites yet reported comparatively lower AUC values. Among the 7 externally validated studies, the reported performance was generally more conservative [6,30,34,35,38,39,44]. Taliaz et al [6] achieved an average balanced accuracy of 70.1% internally, but the only model validated externally yielded 61.3%. Zhukovsky et al [44] similarly demonstrated a performance gap, with the best pretreatment cross-trial AUC of 0.62 to 0.67 for SSRI generalization across 2 independent RCTs. However, Chen et al [39] maintained consistent performance, with R^2 values of 41% to 45% on independent RCT datasets, suggesting that external validation does not uniformly attenuate results when the modeling approach is well aligned to the clinical context. Model architecture, by contrast, appeared to play a comparatively secondary role. Both complex approaches, such as convolutional neural networks, and simpler regularized regression methods achieved strong performance when paired with informative neurobiological features, while ensemble methods applied to clinical-only data did not consistently outperform simpler algorithms [41,43]. This observed pattern indicates that variation in reported performance across studies corresponded more closely to input feature characteristics than to algorithm type.

Risk of Bias Assessment

Risk of bias was assessed separately for model development and validation phases (Multimedia Appendix 2). For model development, the overall risk of bias was low in 10 studies, high in 4 studies, and unclear in 4 studies (Tables S1 and S2 in Multimedia Appendix 2). The participants and predictors domains demonstrated consistently low risk, while the outcome and analysis domains raised more frequent concerns. Common methodological limitations included incomplete reporting of missing data handling, absence of calibration assessment, and insufficient detail regarding overfitting prevention. For model validation, 9 studies were rated as low risk, 5 as high risk, and 5 as unclear. Applicability concerns were low across all studies.

Discussion

Principal Findings

Our review identified only 19 eligible studies out of 5370 initially screened, demonstrating a critical need for research that applies ML for comparative modeling across multiple antidepressant medications. This scarcity represents a critical gap because most studies focused on only a single pharmacological treatment or grouped treatment arms [36,48-57]. These approaches cannot inform the clinician's central

task of selecting between therapeutic options for an individual patient. This field remains constrained by methodological inconsistencies, limited generalizability, and a lack of explainability, which collectively hinder translation into routine clinical practice.

Beyond the scarcity of comparative studies, the uneven distribution of antidepressants examined raises questions about model applicability. Current models are optimized for predicting outcomes among the most commonly prescribed medications. Yet, clinicians often face decisions involving less frequently studied agents, particularly when managing treatment-resistant depression or addressing specific patient contraindications [42]. The scarcity of studies examining tricyclic antidepressants, monoamine oxidase inhibitors, or atypical antidepressants means that ML-guided treatment selection may be least available precisely when clinical decision-making is most complex. Moreover, most studies fail to address combination therapy. Our review found that only 1 study explicitly included patients receiving augmentation therapy and treated all medication combinations as a single category, without distinguishing between specific drug-drug interactions or predicting optimal combination strategies [33]. Modeling each medication separately, or treating all combinations as equivalent, fails to account for the complex pharmacological interactions inherent in combination therapy [58].

Among the included studies, 3 distinct strategies for individualized treatment selection were identified. The first strategy was to train separate models for each antidepressant, which allows for tailored prediction per medication but lacks a unified comparative framework, making it difficult to rank or recommend treatments across options for a single patient. Because each model is trained independently, observed differences in predictive performance may reflect sample variation, modeling noise, or hyperparameter settings, rather than true pharmacological differences. These models, therefore, function primarily as treatment-specific response prediction tools rather than direct treatment recommendation systems.

The second strategy was to cluster patients based on symptom profiles or other baseline characteristics and then assess treatment response within each subgroup [16,30,33,35,39]. These studies aimed to improve treatment matching by identifying homogeneous patient subtypes that might respond differentially to treatment. However, as the number of interventions assessed increases, creating meaningful clusters becomes increasingly difficult due to the exponential growth in possible treatment-subgroup combinations [59]. Additionally, as cluster numbers increase, interpreting which patient characteristics drive differential treatment responses becomes more complex and less clinically actionable [59,60]. This approach relies on unsupervised techniques and post hoc comparisons. It requires large sample sizes to ensure stable clustering solutions. External validation is also needed to confirm that identified subgroups replicate across different populations [61-63]. More generally, this approach supports response stratification rather than comparative treatment selection.

The third strategy focuses on developing a unified differential treatment prediction framework that simultaneously generates calibrated remission probabilities across multiple pharmacological treatments for the same individual patient [42]. This approach directly addresses the comparative selection challenge by producing patient-level treatment rankings rather than relying on post hoc comparisons of separately trained models. Notably, only 1 study implements such a framework, suggesting a need for further investigation [42]. The emergence of this third strategy marks a conceptual shift from “Can we predict response?” toward “Which treatment should this patient receive?” This question most closely mirrors clinical decision-making.

Several studies integrated multimodal data, including neuroimaging, genetic markers, and electroencephalogram signals, generally demonstrating improved performance over clinical-only models [14,29,35,43]. However, this advantage should be interpreted cautiously. Studies incorporating multimodal features simultaneously increased total feature dimensionality, and without systematic ablation studies, it is difficult to disentangle modality-specific signals from the effect of higher data volume [64,65]. This ambiguity is compounded by confounding between data modality, sample size, and validation approach. Neurobiological studies tended to use smaller cohorts with internal validation only, whereas larger studies using clinical data with external validation reported more conservative performance [16,31,37,41]. Practical barriers further limit multimodal integration, as neuroimaging and genetic testing require dedicated infrastructure and personnel, which constrain scalability. Future work should evaluate whether causal inference frameworks designed for heterogeneous treatment effects improve selection accuracy relative to conventional predictive models.

Beyond data modality considerations, a fundamental limitation underlying these strategies is the insufficient distinction between prognostic and predictive features. Prognostic features predict overall treatment outcome regardless of medication choice, whereas predictive features identify differential responses through treatment-by-covariate interactions [66,67]. Only predictive features can directly inform comparative medication selection, but most included studies did not formally distinguish between these 2 types of markers. Among drug-specific modeling studies, differences in model outputs or performance across treatment arms may reflect sampling variability or modeling noise rather than genuine pharmacological specificity [6,14,36,37,41]. Similarly, studies employing clustering, general prediction models, or other frameworks primarily identified prognostic patterns, with post hoc treatment comparisons generating hypotheses rather than confirming predictive effects [16,29-31,33-35,39,40,43]. Two studies from the same research group provided the most direct empirical evidence on this distinction. Iniesta et al [29] demonstrated through cross-drug validation that models trained on escitalopram patients explained negligible variance in nortriptyline outcomes and vice versa. Additionally, in their other study, Iniesta et al [32] extended this finding with a larger genetic feature set,

confirming drug-specific prediction with a validation AUC of 0.77 for both drugs, while cross-drug AUC fell to a chance level between 0.57 and 0.62. However, Zhukovsky et al [44] demonstrated that models trained on sertraline data in one trial could predict escitalopram response in an independent trial at above-chance levels, suggesting that some predictive signal may generalize across SSRIs with similar mechanisms of action. This finding complicates the prognostic-predictive distinction, as it raises the possibility that within-class cross-drug generalizability reflects shared pharmacological pathways rather than purely prognostic features. Future research should prioritize frameworks that explicitly model heterogeneous treatment effects and adopt cross-drug validation to differentiate predictive from prognostic signals.

Another major barrier relates to inconsistencies in outcome definitions and evaluation metrics. Variations in response thresholds and remission criteria can reclassify patients at the margin, potentially influencing model training [68]. Combined with inconsistent evaluation metrics and varying prediction horizons, this heterogeneity hinders reliable cross-study comparison and makes quantitative meta-analysis impossible [69]. To advance the field, consensus development regarding standardized evaluation frameworks will be important.

The near absence of formal explainability techniques across all included studies represents a significant barrier to clinical translation. In psychiatry, explainability is essential for building clinician trust, ensuring ethical accountability, and supporting treatment selection decisions [47,70,71]. A noted barrier to the adoption of ML tools by clinicians relates to predictions without clear explanations, particularly in high-stakes scenarios like antidepressant selection [72]. Future studies should embed explainability into the model development pipeline and collaborate with mental health professionals to ensure clinical relevance.

Dataset redundancy and limited generalizability raise additional concerns. Repeated use of the same datasets introduces risks of overfitting to cohort-specific patterns and model tuning based on familiar distributions, potentially inflating performance estimates [69,73]. The highest reported performances share a common methodological profile. Small sample sizes combined with high-dimensional neurobiological features and internal-only validation, a pattern consistent with overfitting [41]. The performance gap between internal and external validation further supports this concern. For example, Taliaz et al [6] achieved 70.1% balanced accuracy internally but only 61.3% on externally validated models. In addition, Zhukovsky et al [44] demonstrated that models showed reduced performance when applied across independent trials compared to within-trial settings. Moreover, nearly all studies reported only discrimination metrics such as AUC, while calibration, which assesses whether predicted probabilities accurately reflect observed outcomes, was almost entirely absent. Only Benrimoh et al [42] reported calibrated remission probabilities, demonstrating that this approach is feasible within a unified differential prediction framework. For treatment selection, well-calibrated probability estimates are arguably more important than discrimination, as clinicians need reliable absolute risk estimates to compare treatment alternatives [74-76]. Limited geographic and demographic diversity compounds these concerns, as genetic variation and cultural differences in symptom expression may cause models to over-rely on region-specific features [36,77-80]. The predominance of internal validation, sometimes using overlapping cohorts as ostensibly independent test sets, further limits confidence in reported performance. Future studies should prioritize external validation on prospective, multisite, demographically diverse cohorts and report calibration alongside discrimination metrics.

Finally, few studies have addressed long-term outcomes. Most predictions have targeted short-term responses within 6 to 12 weeks, with very limited attention given to sustained remission, relapse prevention, or longitudinal treatment trajectories. However, prediction horizons inherently differ in difficulty. Longer-term forecasting is more challenging due to increased variability, patient dropout, and cumulative confounders [81]. Moreover, prediction horizons correspond to distinct clinical applications. Short-term response prediction can guide early treatment adjustments, while long-term modeling is crucial for evaluating real-world effectiveness and preventing relapses [82]. Yet these long-term outcomes are central to real-world decision-making and should be prioritized in future modeling efforts.

Limitations

We acknowledge the limitations in our review. First, our search was restricted to 3 databases and English-language publications, potentially missing relevant studies in other databases, languages, or gray literature sources. Second, we may have missed relevant studies due to variations in indexing terminology across the interdisciplinary field of ML and psychiatry, where different vocabularies are used to describe similar comparative treatment prediction approaches. Third, although the risk of bias was assessed using PROBAST-AI, all 19 studies were retained regardless of their bias classification due to the limited number of eligible studies. Findings from studies rated as high risk should therefore be interpreted with caution. Fourth, the heterogeneity in study designs, methodologies, and reporting standards across included studies prevented quantitative meta-analysis, limiting our review to a narrative synthesis and reducing our ability to provide pooled estimates of model performance. Finally, the confounding between data modality, sample size, and validation approach across included studies prevents us from isolating the independent contribution of multimodal features to model performance, and future ablation studies are needed to clarify whether observed performance advantages reflect modality-specific information or greater overall data volume.

Conclusion

This systematic review reveals that ML for comparative antidepressant selection remains in an early stage of development. Three distinct modeling strategies were identified, which are drug-specific models, subtype- or trajectory-based approaches, and a unified differential

prediction framework. Yet only the last directly supports patient-level treatment ranking, and it was implemented in a single study. Most studies did not formally distinguish between prognostic and predictive features, limiting the capacity to identify which patients benefit differentially from specific medications. Additional barriers include limited cross-trial validation, near-absent calibration reporting, and the lack of formal explainability techniques. To advance toward clinical translation, future research should prioritize unified comparative frameworks with calibrated predictions, rigorous external validation on diverse cohorts, explicit modeling of heterogeneous treatment effects, and integration of explainability into model development. Only by addressing these fundamental gaps can ML become a trustworthy and clinically valuable tool for optimizing antidepressant selection in MDD.

Acknowledgments

The authors extend their gratitude to Evan Sterling and Lisa Shin (Morisset Library, University of Ottawa) for their invaluable support in guiding our research methodology, which has been crucial for this systematic review. Google Translate and Grammarly were used only for translation assistance and language polishing. They were not used for scientific content generation, data analysis, result interpretation, or reference creation. The authors reviewed all revisions and take full responsibility for the manuscript.

Funding

The authors declare that no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy.
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Risk of bias assessment.
[DOCX File (Microsoft Word File), 50 KB-Multimedia Appendix 2]

Checklist 1

PRISMA checklist.
[DOCX File (Microsoft Word File), 14 KB-Checklist 1]

References

1. Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nat Rev Dis Primers. Sep 15, 2016;2(1):16065. [doi: 10.1038/nrdp.2016.65] [Medline: 27629598]
2. Christensen MC, Wong CMJ, Baune BT. Symptoms of major depressive disorder and their impact on psychosocial functioning in the different phases of the disease: do the perspectives of patients and healthcare providers differ? Front Psychiatry. 2020;11:280. [doi: 10.3389/fpsyt.2020.00280] [Medline: 32390877]
3. Greer TL, Kurian BT, Trivedi MH. Defining and measuring functional recovery from depression. CNS Drugs. Apr 2010;24(4):267-284. [doi: 10.2165/11530230-000000000-00000] [Medline: 20297853]
4. Cai H, Xie XM, Zhang Q, et al. Prevalence of suicidality in major depressive disorder: a systematic review and meta-analysis of comparative studies. Front Psychiatry. 2021;12:690130. [doi: 10.3389/fpsyt.2021.690130] [Medline: 34603096]
5. Goldberg D. The heterogeneity of "major depression". World Psychiatry. Oct 2011;10(3):226-228. [doi: 10.1002/j.2051-5545.2011.tb00061.x] [Medline: 21991283]
6. Taliaz D, Spinrad A, Barzilay R, et al. Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data. Transl Psychiatry. Jul 8, 2021;11(1):381. [doi: 10.1038/s41398-021-01488-3] [Medline: 34238923]
7. Cipriani A, Salanti G, Furukawa TA, et al. Antidepressants might work for people with major depression: where do we go from here? Lancet Psychiatry. Jun 2018;5(6):461-463. [doi: 10.1016/S2215-0366(18)30133-0] [Medline: 29628364]
8. Culpepper L, Martin A, Nabulsi N, Parikh M. The humanistic and economic burden associated with major depressive disorder: a retrospective cross-sectional analysis. Adv Ther. May 2024;41(5):1860-1884. [doi: 10.1007/s12325-024-02817-w] [Medline: 38466558]
9. Citrome L, Jain R, Tung A, Landsman-Blumberg PB, Kramer K, Ali S. Prevalence, treatment patterns, and stay characteristics associated with hospitalizations for major depressive disorder. J Affect Disord. Apr 15, 2019;249:378-384. [doi: 10.1016/j.jad.2019.01.044] [Medline: 30818246]

10. Sheu YH, Magdamo C, Miller M, Das S, Blacker D, Smoller JW. AI-assisted prediction of differential response to antidepressant classes using electronic health records. NPJ Digit Med. Apr 26, 2023;6(1):73. [doi: 10.1038/s41746-023-00817-8] [Medline: 37100858]
11. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform. Aug 2021;25(8):3121-3129. [doi: 10.1109/JBHI.2021.3063721] [Medline: 33661740]
12. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: a scoping review. NPJ Digit Med. Nov 7, 2022;5(1):171. [doi: 10.1038/s41746-022-00712-8] [Medline: 36344814]
13. Pei C, Sun Y, Zhu J, et al. Ensemble learning for early-response prediction of antidepressant treatment in major depressive disorder. J Magn Reson Imaging. Jul 2020;52(1):161-171. [doi: 10.1002/jmri.27029] [Medline: 31859419]
14. Curtiss J, Smoller JW, Pedrelli P. Optimizing precision medicine for second-step depression treatment: a machine learning approach. Psychol Med. Jul 2024;54(10):2361-2368. [doi: 10.1017/S0033291724000497] [Medline: 38533794]
15. Joyce JB, Grant CW, Liu D, et al. Multi-omics driven predictions of response to acute phase combination antidepressant therapy: a machine learning approach with cross-trial replication. Transl Psychiatry. Oct 7, 2021;11(1):513. [doi: 10.1038/s41398-021-01632-z] [Medline: 34620827]
16. Fu CHY, Antoniades M, Erus G, et al. Neuroanatomical dimensions in medication-free individuals with major depressive disorder and treatment response to SSRI antidepressant medications or placebo. Nat Ment Health. 2024;2(2):164-176. [doi: 10.1038/s44220-023-00187-w] [Medline: 38948238]
17. Grzenda A, Speier W, Siddarth P, et al. Machine learning prediction of treatment outcome in late-life depression. Front Psychiatry. 2021;12:738494. [doi: 10.3389/fpsyt.2021.738494] [Medline: 34744829]
18. Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. Jul 22, 2010;363(4):301-304. [doi: 10.1056/NEJMp1006304] [Medline: 20551152]
19. Marx W, Penninx BWJH, Solmi M, et al. Major depressive disorder. Nat Rev Dis Primers. Aug 24, 2023;9(1):44. [doi: 10.1038/s41572-023-00454-1] [Medline: 37620370]
20. Cui L, Li S, Wang S, et al. Major depressive disorder: hypothesis, mechanism, prevention and treatment. Signal Transduct Target Ther. Feb 2024;9(1):30. [doi: 10.1038/s41392-024-01738-y] [Medline: 38331979]
21. Rush AJ, Trivedi MH, Wisniewski SR, et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. N Engl J Med. Mar 23, 2006;354(12):1231-1242. [doi: 10.1056/NEJMoa052963] [Medline: 16554525]
22. Papakostas GI, Fava M, Thase ME. Treatment of SSRI-resistant depression: a meta-analysis comparing within- versus across-class switches. Biol Psychiatry. Apr 2008;63(7):699-704. [doi: 10.1016/j.biopsych.2007.08.010] [Medline: 17919460]
23. Mehltretter J, Fratila R, Benrimoh DA, et al. Differential treatment benet prediction for treatment selection in depression: a deep learning analysis of STAR*D and CO-MED data. Comput Psychiatr. 2020;4:61. [doi: 10.1162/cpsy_a_00029]
24. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”. Intell Med. Feb 2024;4(1):52-57. [doi: 10.1016/j.imed.2023.08.001]
25. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. Jul 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. Jan 1, 2019;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
28. 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]
29. Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. Jul 2016;78:94-102. [doi: 10.1016/j.jpsychires.2016.03.016] [Medline: 27089522]
30. Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry. Apr 1, 2017;74(4):370-378. [doi: 10.1001/jamapsychiatry.2017.0025] [Medline: 28241180]
31. Crane NA, Jenkins LM, Bhaumik R, et al. Multidimensional prediction of treatment response to antidepressants with cognitive control and functional MRI. Brain. Feb 2017;140(2):472-486. [doi: 10.1093/brain/aww326] [Medline: 28122876]
32. Iniesta R, Hodgson K, Stahl D, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep. Apr 3, 2018;8(1):5530. [doi: 10.1038/s41598-018-23584-z] [Medline: 29615645]

33. Kautzky A, Möller HJ, Dold M, et al. Combining machine learning algorithms for prediction of antidepressant treatment response. Acta Psychiatr Scand. Jan 2021;143(1):36-49. [doi: 10.1111/acps.13250] [Medline: 33141944]
34. Hughes MC, Pradier MF, Ross AS, McCoy TH Jr, Perlis RH, Doshi-Velez F. Assessment of a prediction model for antidepressant treatment stability using supervised topic models. JAMA Netw Open. May 1, 2020;3(5):e205308. [doi: 10.1001/jamanetworkopen.2020.5308] [Medline: 32432711]
35. Athreya AP, Brückl T, Binder EB, et al. Prediction of short-term antidepressant response using probabilistic graphical models with replication across multiple drugs and treatment settings. Neuropsychopharmacol. Jun 2021;46(7):1272-1282. [doi: 10.1038/s41386-020-00943-x] [Medline: 33452433]
36. Bi Y, Ren D, Guo Z, et al. Influence and interaction of genetic, cognitive, neuroendocrine and personalistic markers to antidepressant response in Chinese patients with major depression. Prog Neuropsychopharmacol Biol Psychiatry. Jan 10, 2021;104:110036. [doi: 10.1016/j.pnpbp.2020.110036] [Medline: 32702381]
37. Nguyen KP, Chin Fatt C, Treacher A, et al. Patterns of pretreatment reward task brain activation predict individual antidepressant response: key results from the EMBARC randomized clinical trial. Biol Psychiatry. Mar 15, 2022;91(6):550-560. [doi: 10.1016/j.biopsych.2021.09.011] [Medline: 34916068]
38. Wang X, Qin J, Zhu R, et al. Predicting treatment selections for individuals with major depressive disorder according to functional connectivity subgroups. Brain Connect. Oct 2022;12(8):699-710. [doi: 10.1089/brain.2021.0153] [Medline: 34913731]
39. Chen Y, Stewart JW, Ge J, Cheng B, Chekroud A, Hellerstein DJ. Personalized symptom clusters that predict depression treatment outcomes: a replication of machine learning methods. J Affect Disord Rep. Jan 2023;11:100470. [doi: 10.1016/j.jadr.2023.100470]
40. Turner RJ, Hagoort K, Meijer RJ, Coenen F, Scheepers FE. Bayesian network analysis of antidepressant treatment trajectories. Sci Rep. May 24, 2023;13(1):8428. [doi: 10.1038/s41598-023-35508-7] [Medline: 37225783]
41. Ravan M, Noroozi A, Gediya H, James Basco K, Hasey G. Using deep learning and pretreatment EEG to predict response to sertraline, bupropion, and placebo. Clin Neurophysiol. Nov 2024;167:198-208. [doi: 10.1016/j.clinph.2024.09.002] [Medline: 39332081]
42. Benrimoh D, Armstrong C, Mehltretter J, et al. Development of the treatment prediction model in the artificial intelligence in depression - medication enhancement study. Npj Mental Health Res. Jun 2025;4(1):26. [doi: 10.1038/s44184-025-00136-8] [Medline: 40550942]
43. Carr E, Rietschel M, Mors O, et al. Optimizing the prediction of depression remission: a longitudinal machine learning approach. Am J Med Genet B Neuropsychiatr Genet. Apr 2025;198(3):e33014. [doi: 10.1002/ajmg.b.33014] [Medline: 39470297]
44. Zhukovsky P, Trivedi MH, Weissman M, Parsey R, Kennedy S, Pizzagalli DA. Generalizability of treatment outcome prediction across antidepressant treatment trials in depression. JAMA Netw Open. Mar 3, 2025;8(3):e251310. [doi: 10.1001/jamanetworkopen.2025.1310] [Medline: 40111362]
45. Bing maps. Microsoft | Marketplace. URL: https://marketplace.microsoft.com/en-au/product/office/WA102957661?tab=Overview [Accessed 2026-05-07]
46. Microsoft bing maps platform APIs terms of use. Bing maps | Dev Center. URL: https://www.bingmapsportal.com/terms [Accessed 2026-05-07]
47. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med. Jan 18, 2023;6(1):6. [doi: 10.1038/s41746-023-00751-9] [Medline: 36653524]
48. Ho CSH, Wang J, Tay GWN, et al. Application of functional near-infrared spectroscopy and machine learning to predict treatment response after six months in major depressive disorder. Transl Psychiatry. Jan 11, 2025;15(1):7. [doi: 10.1038/s41398-025-03224-7] [Medline: 39799114]
49. Gomeni R, Bressolle-Gomeni F. Comparison of different machine learning methodologies for predicting the non-specific treatment response in placebo controlled major depressive disorder clinical trials. Clin Transl Sci. Jan 2025;18(1):e70128. [doi: 10.1111/cts.70128] [Medline: 39807769]
50. Bossarte RM, Ross EL, Liu H, et al. Development of a model to predict combined antidepressant medication and psychotherapy treatment response for depression among veterans. J Affect Disord. Apr 1, 2023;326:111-119. [doi: 10.1016/j.jad.2023.01.082] [Medline: 36709831]
51. Deco G, Sanz Perl Y, Johnson S, Bourke N, Carhart-Harris RL, Kringelbach ML. Different hierarchical reconfigurations in the brain by psilocybin and escitalopram for depression. Nat Mental Health. Sep 2024;2(9):1096-1110. [doi: 10.1038/s44220-024-00298-y]
52. Carrillo F, Sigman M, Fernández Slezak D, et al. Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. J Affect Disord. Apr 1, 2018;230:84-86. [doi: 10.1016/j.jad.2018.01.006] [Medline: 29407543]

53. Copa D, Erritzoe D, Giribaldi B, Nutt D, Carhart-Harris R, Tagliazucchi E. Predicting the outcome of psilocybin treatment for depression from baseline fMRI functional connectivity. J Affect Disord. May 15, 2024;353:60-69. [doi: 10.1016/j.jad.2024.02.089] [Medline: 38423367]
54. Browning M, Kingslake J, Dourish CT, Goodwin GM, Harmer CJ, Dawson GR. Predicting treatment response to antidepressant medication using early changes in emotional processing. Eur Neuropsychopharmacol. Jan 2019;29(1):66-75. [doi: 10.1016/j.euroneuro.2018.11.1102] [Medline: 30473402]
55. Bao Z, Zhao X, Li J, et al. Prediction of repeated-dose intravenous ketamine response in major depressive disorder using the GWAS-based machine learning approach. J Psychiatr Res. Jun 2021;138:284-290. [doi: 10.1016/j.jpsychires.2021.04.014] [Medline: 33878621]
56. Davey CG, Cearns M, Jamieson A, Harrison BJ. Suppressed activity of the rostral anterior cingulate cortex as a biomarker for depression remission. Psychol Med. Apr 2023;53(6):2448-2455. [doi: 10.1017/S0033291721004323] [Medline: 36762975]
57. Banerjee S, Wu Y, Bingham KS, et al. Trajectories of remitted psychotic depression: identification of predictors of worsening by machine learning. Psychol Med. Apr 2024;54(6):1142-1151. [doi: 10.1017/S0033291723002945] [Medline: 37818656]
58. Nagy T, Gonda X, Gezsi A, et al. Pharmacological profiling of major depressive disorder-related multimorbidity clusters. Eur Neuropsychopharmacol. Jul 2025;96:71-83. [doi: 10.1016/j.euroneuro.2025.05.007] [Medline: 40483774]
59. Datta E, Ballal A, López JE, Izu LT. MapperPlus: agnostic clustering of high-dimension data for precision medicine. PLoS Digit Health. Aug 2023;2(8):e0000307. [doi: 10.1371/journal.pdig.0000307] [Medline: 37556425]
60. Kent P, Stochkendahl MJ, Christensen HW, Kongsted A. Could the clinical interpretability of subgroups detected using clustering methods be improved by using a novel two-stage approach? Chiropr Man Therap. 2015;23(1):20. [doi: 10.1186/s12998-015-0064-9] [Medline: 26140192]
61. Benrimoh D, Kleinerman A, Furukawa TA, et al. Towards outcome-driven patient subgroups: a machine learning analysis across six depression treatment studies. Am J Geriatr Psychiatry. Mar 2024;32(3):280-292. [doi: 10.1016/j.jagp.2023.09.009] [Medline: 37839909]
62. Kim IB, Park SC. Machine learning-based definition of symptom clusters and selection of antidepressants for depressive syndrome. Diagnostics (Basel). Sep 7, 2021;11(9):1631. [doi: 10.3390/diagnostics11091631] [Medline: 34573974]
63. Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of clustering multimodal clinical data: review of applications in asthma subtyping. JMIR Med Inform. May 28, 2020;8(5):e16452. [doi: 10.2196/16452] [Medline: 32463370]
64. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. Feb 2019;41(2):423-443. [doi: 10.1109/TPAMI.2018.2798607] [Medline: 29994351]
65. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. Oct 15, 2018;180(Pt A):68-77. [doi: 10.1016/j.neuroimage.2017.06.061] [Medline: 28655633]
66. Lyons M, Delgadillo J. A systematic review of predictors and moderators of treatment response in psychological interventions for persisting forms of depression. Br J Clin Psychol. Sep 2025;64(3):623-656. [doi: 10.1111/bjc.12513] [Medline: 39737557]
67. Fennema D, Barker GJ, O'Daly O, et al. Neural signatures of emotional biases predict clinical outcomes in difficult-to-treat depression. Res Dir Depress. 2024;1:e21. [doi: 10.1017/dep.2024.6] [Medline: 40028885]
68. Lisinski A, Hieronymus F, Nilsson S, Eriksson E. Impact of chosen cutoff on response rate differences between selective serotonin reuptake inhibitors and placebo. Transl Psychiatry. Apr 14, 2022;12(1):160. [doi: 10.1038/s41398-022-01882-5] [Medline: 35422023]
69. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R. Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol Artif Intell. Jan 2023;5(1):e220028. [doi: 10.1148/ryai.220028] [Medline: 36721408]
70. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. Jan 2021;113:103655. [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
71. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. Proc 4th Machine Learning Healthcare Conference. 2019;106:359-380. URL: https://proceedings.mlr.press/v106/tonekaboni19a.html [Accessed 2026-04-29]
72. Bradshaw TJ, McCradden MD, Jha AK, et al. Artificial intelligence algorithms need to be explainable-or do they? J Nucl Med. Jun 2023;64(6):976-977. [doi: 10.2967/jnumed.122.264949] [Medline: 37116913]
73. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: data, validation, communication challenges. Med Phys. Oct 2018;45(10):e834-e840. [doi: 10.1002/mp.12811] [Medline: 30144098]

74. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. Jan 2010;21(1):128-138. [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
75. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. Sep 3, 2024;386:e078276. [doi: 10.1136/bmj-2023-078276] [Medline: 39227063]
76. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. Jan 15, 2024;384:e074820. [doi: 10.1136/bmj-2023-074820] [Medline: 38224968]
77. Fratelli C, Siqueira J, Silva C, Ferreira E, Silva I. 5HTTLPR genetic variant and major depressive disorder: a review. Genes (Basel). Oct 26, 2020;11(11):1260. [doi: 10.3390/genes11111260] [Medline: 33114535]
78. Goldman N, Glei DA, Lin YH, Weinstein M. The serotonin transporter polymorphism (5-HTTLPR): allelic variation and links with depressive symptoms. Depress Anxiety. Mar 2010;27(3):260-269. [doi: 10.1002/da.20660] [Medline: 20196101]
79. Dere J, Sun J, Zhao Y, et al. Beyond "somatization" and "psychologization": symptom-level variation in depressed Han Chinese and Euro-Canadian outpatients. Front Psychol. 2013;4:377. [doi: 10.3389/fpsyg.2013.00377] [Medline: 23818884]
80. Choi E, Chentsova-Dutton Y, Parrott WG. The effectiveness of somatization in communicating distress in Korean and American cultural contexts. Front Psychol. 2016;7:383. [doi: 10.3389/fpsyg.2016.00383] [Medline: 27047414]
81. Cascarano A, Mur-Petit J, Hernández-González J, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif Intell Rev. Nov 2023;56(S2):1711-1771. [doi: 10.1007/s10462-023-10561-w]
82. 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]

Abbreviations

AUC: area under the curve
MDD: major depressive disorder
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST-AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence
RCT: randomized controlled trial
SSRI: Selective Serotonin Reuptake Inhibitor

Edited by John Torous; peer-reviewed by Alec J Jamieson, John-Jose Nunez; submitted 12.Dec.2025; final revised version received 16.Mar.2026; accepted 17.Mar.2026; published 13.May.2026

Please cite as:
He F, Huang S, Wang R, Chang A, Phillips JL, Sun C
Machine Learning for Comparative Antidepressant Selection in Major Depressive Disorder: Systematic Review
JMIR Ment Health 2026;13:e89352
URL: https://mental.jmir.org/2026/1/e89352
doi: 10.2196/89352