Taylor & Francis
Taylor & Francis Group

Expert Review of Clinical Pharmacology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ierj20

Machine learning, pharmacogenomics, and clinical psychiatry: predicting antidepressant response in patients with major depressive disorder

William V. Bobo, Bailey Van Ommeren & Arjun P. Athreya

To cite this article: William V. Bobo, Bailey Van Ommeren & Arjun P. Athreya (2022) Machine learning, pharmacogenomics, and clinical psychiatry: predicting antidepressant response in patients with major depressive disorder, Expert Review of Clinical Pharmacology, 15:8, 927-944, DOI: 10.1080/17512433.2022.2112949

To link to this article: https://doi.org/10.1080/17512433.2022.2112949

EXPERT REVIEW OF CLINICAL PHARMACOLOGY
2022, VOL. 15, NO. 8, 927–944
https://doi.org/10.1080/17512433.2022.2112949

Taylor & Francis
Taylor & Francis Group

REVIEW

Machine learning, pharmacogenomics, and clinical psychiatry: predicting antidepressant response in patients with major depressive disorder

William V. Bobo[a,b], Bailey Van Ommeren[c] and Arjun P. Athreya[d]

[a]Department of Psychiatry & Psychology, Mayo Clinic Florida, Jacksonville, FL, USA; [b]Center for Individualized Medicine, Mayo Clinic, Rochester, FL, USA; [c]Department of Research and Education, Mayo Clinic, Rochester, MN, USA; [d]Department of Molecular Pharmacology & Experimental Therapeutics, Mayo Clinic, Rochester, MN, USA

ABSTRACT

Introduction: The efficacy of antidepressants for patients with major depressive disorder (MDD) varies from individual to individual, making the prediction of therapeutic outcomes difficult. Better methods for predicting antidepressant outcomes are needed. However, complex interactions between biological, psychological, and environmental factors affect outcomes, presenting immense computational challenges for prediction. Using machine learning (ML) techniques with pharmacogenomics data provides one pathway toward individualized prediction of therapeutic outcomes of antidepressants.

Areas covered: This report systematically reviews the methods, results, and limitations of individual studies of ML and pharmacogenomics for predicting response and/or remission with antidepressants in patients with MDD. Future directions for research and pragmatic considerations for the clinical implementation of ML-based pharmacogenomic algorithms are also discussed.

Expert opinion: ML methods utilizing pharmacogenomic and clinical data demonstrate promising results for predicting short-term antidepressant response. However, predictions of antidepressant treatment outcomes depend on contextual factors that ML algorithms may not be able to capture. As such, ML-driven prediction is best viewed as a companion to clinical judgment, not its replacement. Successful implementation and adoption of methods predicting antidepressant response warrants provider education about ML and close collaborations between computing scientists, pharmacogenomic experts, health system engineers, laboratory medicine experts, and clinicians.

ARTICLE HISTORY
Received 18 April 2022
Accepted 9 August 2022

KEYWORDS
Machine learning; artificial intelligence; deep learning; genomics; pharmacogenomics; prediction; major depressive disorder; depression; antidepressant; outcome

1. Introduction

Major depressive disorder (MDD) affects over 264 million people worldwide [1], making it one of the most prevalent illnesses in medicine [2]. MDD is traditionally described as an episodic illness; however, many patients continue to experience persisting symptoms between syndromal relapses that are associated with poor quality of life and functioning in nearly every domain [3,4]. Depression is considered the leading cause of disability associated with chronic illness worldwide and is a leading cause of early mortality due to general medical illness and suicide [5–7]. Not surprisingly, the societal costs associated with MDD are staggering [8], totaling over $210 billion USD in 2010 [9].

For many patients, the symptoms of MDD can be managed with evidence-based psychosocial treatment and appropriate pharmacotherapy. Unfortunately, only one-third of depressed patients who receive antidepressants achieves remission [10,11], which is considered the goal of treatment by both clinicians and patients [12]. Additionally, multiple therapeutic trials, each lasting several weeks, are often required before achieving a good outcome from treatment [13]. Given the lack of a robust evidence base for selecting initial and next-step antidepressants for depressed patients, the pharmacological treatment of depression often resembles an 'artisanal' or 'try-and-try-again' approach [14]. That is, treatment selection and management are based mainly on intuition and experience rather than quantitative predictive factors that serve as a companion to clinical judgment.

These challenges highlight the importance of developing better methods for predicting outcomes of treatment with a given antidepressant based on an individual patient's unique biological and clinical characteristics. Decades of research has identified clinical predictors of poor response to selective serotonin reuptake inhibitors (SSRIs) and other antidepressants, but, with few exceptions, these are only minimally predictive of outcomes [15]. Historically, better responses to certain types of antidepressants have been suggested to occur for specific clinical subtypes of depressed patients [16,17]. However, these results have been difficult to replicate due, in part, to the considerable overlap between depressive subtypes [18]. Similarly, symptom clustering approaches have identified groups of patients with differential antidepressant response trajectories, but they are less useful for predicting discrete treatment outcomes at the individual patient level. Although a personalized treatment approach is desirable, no single set of assessments can yet predict antidepressant outcome with sufficient validity for clinical use [19].

In the last decade, an increasing body of research has shown that integrating pharmacogenomic markers of

CONTACT William V. Bobo bobo.william@mayo.edu Department of Psychiatry & Psychology, Mayo Clinic Florida, Jacksonville, FL, USA

Article highlights

- Pharmacogenomic data have been used to predict short-term clinical responses to treatment with antidepressants in people with depression using a variety of machine learning methods.
- The results of most studies show that high and generally comparable levels of predictive performance can be achieved using these methods; however, individual studies vary widely regarding the machine learning methods, pharmacogenomic features, non-pharmacogenomic features, validation methods, and study drugs that were used, making direct comparisons between the reviewed studies difficult to conduct.
- Few studies included an independent dataset, separate from the original dataset(s) used for algorithm development, for validation of algorithm performance.
- Several factors may limit both the validity and clinical utility of the predictions achieved by machine learning models for the treatment of depression with antidepressants, including hidden biases in the data, unpredictable transformations of the data within the algorithms themselves, and the inability of machine learning algorithms to consider important but 'unseen' factors that are not specifically accounted for in the input data.
- For the treatment of depression with antidepressants, machine learning-based tools may be best viewed as companions to clinical judgment within a shared decision-making framework, as opposed to being a driver of clinical decisions.

response to antidepressants with machine learning prediction models may lead, in some cases, to robust predictions of therapeutic outcome [20]. Such approaches constitute an important step toward achieving the goal of individualized treatment selection of antidepressants in depressed patients. Here, we systematically review published studies focused on the integration of machine learning algorithms and pharmacogenomics for purposes of predicting the response to antidepressants in people with MDD. Directions for future research and integration into practice are also discussed.

2. Predicting response to antidepressants in depressed patients: the problem of heterogeneity

The ability to develop reliable (replicable) and valid models for predicting therapeutic responses to antidepressants in depressed patients is limited by several factors, including heterogeneity in disease manifestation and treatment response. Like most psychiatric disorders, MDD is a complex phenotype that is almost certainly not the result of a single etiological factor [21]. The clinical diagnosis of MDD is derived from a set of symptoms that, together, are required to meet diagnostic criteria [22]. There are more than 220 combinations of depressive symptoms [23], each with their own biological foundations and psychosocial interactions [24,25], which can lead to the diagnostic criteria for MDD being met. The severity of individual depressive symptoms can differ widely between patients who meet the same diagnostic criteria for MDD [26], adding an additional layer of complexity in disease presentation. Not surprisingly, there are hundreds – if not thousands – of ways in which individual depressive symptoms can change over time after the initiation of antidepressant treatment in patients who all share the same clinical diagnosis, even with only short-term follow-up [27]. These sources of heterogeneity and the mélange of inter-weaving biological, psychological, and social/environmental factors that are likely underlying pose significant challenges for achieving – let along predicting – antidepressant treatment response in one group of patients and replicating those results in independent groups (or datasets) [26,28].

3. Machine learning, statistical learning, and pharmacogenomics

Heterogeneity in disease manifestation and treatment response creates immense computational challenges for achieving reliable and valid prediction of outcomes with antidepressant treatment and other phenotypes within psychiatry [29]. Consequently, the application of machine learning techniques for predicting the response to treatment with antidepressants has gained traction as a compliment to traditional regression-based statistical approaches [20]. Although the terms 'machine learning' and 'artificial intelligence' are often used interchangeably, machine learning may be best regarded as a set of methodologies under the broader umbrella of artificial intelligence that seek to learn patterns associated with types within a given dataset. Unsupervised machine learning methods use data without labels to infer subtypes (clusters) based on statistical properties (e.g. Euclidean distances) [30]. Supervised machine learning methods use labeled data (e.g. treatment outcomes) from a 'training' cohort to derive models for predicting labels in separate 'testing/validation' (with no overlap from training samples) cohort [31]. Several common (if not exhaustive) supervised machine learning techniques are summarized in Table 1 [31–33].

Previous studies have shown that machine learning approaches that used sociodemographic and clinical measures to predict response to various antidepressants performed significantly better than chance, with areas under the receiver operating curve (AUCs) falling generally in the range of 0.54–0.67 [34–37]. Although some classification models using these predictors yielded higher accuracies for predicting antidepressant treatment outcome phenotypes [38], clinical and sociodemographic measures alone have not generally proven to be sufficiently useful as clinical outcome predictors for individualizing treatment decisions for patients with MDD.

To enable individualized antidepressant treatment selection, analytic approaches that can integrate clinical measures with multiple types of biological response predictors at the individual patient level are needed [37]. The utilization and continuous refinement of predictive approaches utilizing -omic biomarkers (e.g. genomics, metabolomics, and proteomics) are of interest given the increasing knowledge of their associations with mechanisms of antidepressant drug response or MDD pathophysiology or both. Furthermore, advances in high-throughput biological assays are enabling researchers to generate large-scale data at ever-decreasing costs to researchers and health-care systems [39]. These scientific advances will facilitate the opportunity of developing laboratory-based biomarker panels that can be augmented with clinical measures for predicting the clinical effects of antidepressants and individualizing antidepressant treatment selection, even if relatively few predictive biomarkers have made their way into routine clinical practice today [19].

Table 1. Supervised machine and statistical learning approaches and their potential relevance to the prediction of response and remission in people with major depressive disorder who are treated with antidepressants.

Machine learning approach	Definition	Potential relevance for predicting antidepressant response or remission
Supervised learning methods		
Decision tree (DT) methods	Non-parametric approaches that have a flowchart-like appearance whereby data are continuously split according to specified parameters to perform a prediction task.	DT generally has good interpretability but is prone to over-fitting. Tree pruning heuristics are needed to generate compact trees.
Gradient boosting machines (GBM)	An approach whereby new DT models are fit consecutively; errors from 'earlier' trees are used to improve predictions in the 'subsequent' trees.	Theoretically designed to achieve near perfect predictions in training samples. Extensive cross-validation required to learn optimal model parameters. Since multiple decision trees are used for prediction, GBMs are considered a subtype of ensemble machine learning as well as a subtype of DT methods.
Random forests (RF)	An approach that creates several decision trees that, together, are used to perform a prediction task.	RFs offer good bias-variance tradeoffs and are less prone to overfit. Since multiple decision trees are used for prediction, RF is considered a subtype of ensemble machine learning as well as a subtype of DT methods.
Deep learning	Algorithms that learn patterns or data representation through the construction of large neural networks (algorithms that work in multiple layers). Neural networks are often used for unsupervised learning tasks as well as supervised learning tasks.	Deep learning methods can derive predictions from large volumes of complex data and can learn on its own but requires high computing power and may still rely on significant experimenter input to select optimal parameters for the learning algorithm for best predictive performance. Deep learning models offer little interpretability of the results or predictors.
Ensemble machine learning	An approach that creates or combines multiple base models to produce a single optimal model to yield more accurate predictions than would be possible with a single base model.	Provides the ability to overcome the biases of individual learners and reduce error rates based on aggregated results across multiple learners. Computationally expensive due to the need to learn and optimize multiple models.
K-nearest neighbor (kNN)	A non-parametric approach to classification whereby a test object (data point) is compared to the other data points that are most proximal to determine its classification.	Easy to implement and fast learner based on distance and number of neighbors. May not always be sensitive to outliers, high-dimensional feature space, and imbalanced data.
Logistic regression with elastic net penalty (EN)	Logistic regression with regularization to avoid overfit – in this case, by intentionally biasing the data by adding penalties equal to the absolute value and square of the magnitude of regression coefficients.	Easy to implement algorithms with interpretable model parameters. Prediction performance is high for linearly separable data.
Logistic regression with l2 penalty	Logistic regression with regularization to avoid overfit – in this case, by intentionally biasing the data by adding a penalty equal to the square of the magnitude of regression coefficients.	
Support vector machine (SVM)	A non-parametric approach to classification that separates complex observations (data points) by identifying a hyperplane that optimally divides these observations into classes.	Provides the flexibility to achieve predictions in higher-dimensions wherein samples classes could be separated by linear or non-linear planes.

Response to antidepressants is partially influenced by heritable factors [40]. Pharmacogenomics is the study of the contribution of genomics to variation in drug response phenotypes. Therefore, pharmacogenomics has been an essential discipline in the field's attempts to identify biomarkers and associated mechanisms that are capable of distinguishing depressed patients who respond positively or poorly to treatment [41,42]. Although not all studies are in agreement [43,44], pharmacogenomic tailoring of antidepressant selection has shown promise for improving treatment outcomes for antidepressant-treated patients with MDD [45–47]. When combined with machine learning, the complex interactions between genetic variants, non-pharmacogenomic biomarkers, clinical measures, and sociodemographic characteristics may be identified (learned) and, if validated, may be exploited for purposes of response prediction in real-world practice [32].

4. Integration of machine learning and pharmacogenomics: methodology and review of the evidence

4.1. Methodology

We conducted a Medical Subject Heading (MeSH) search in PubMed using the following strategy: (major depressive disorder OR depression*) AND (antidepressants) AND (artificial intelligence OR machine learning OR deep learning OR statistical learning OR prediction) AND (response OR outcome) AND (genomics* OR pharmacogenomics* OR genetics*). The initial search yielded 482 reports. This initial list was narrowed to the 16 reports reviewed below, after excluding articles that did not present the results of original research (e.g. systematic reviews, commentaries, perspectives, and opinions); did not include machine learning or AI methods; were unrelated to antidepressant treatment of MDD; and were published only in abstract or protocol form (Figure 1). WVB and APA screened the articles independently, and a consensus list was derived by adjudicating on any differences in the classification of articles to be included.

4.2. Machine learning and pharmacogenomics for antidepressant response prediction: general overview

A variety of machine learning methods have been applied to longitudinal clinical datasets that generally include a rich array of clinical and sociodemographic variables and pharmacogenomic markers (Table 2). Despite the variety in machine learning approaches across individual studies, all studies involved cohorts of patients with MDD who received short-term treatment with antidepressants and followed a similar general approach (Figure 2). Genomic SNPs used for the prediction of antidepressant treatment responses were either selected

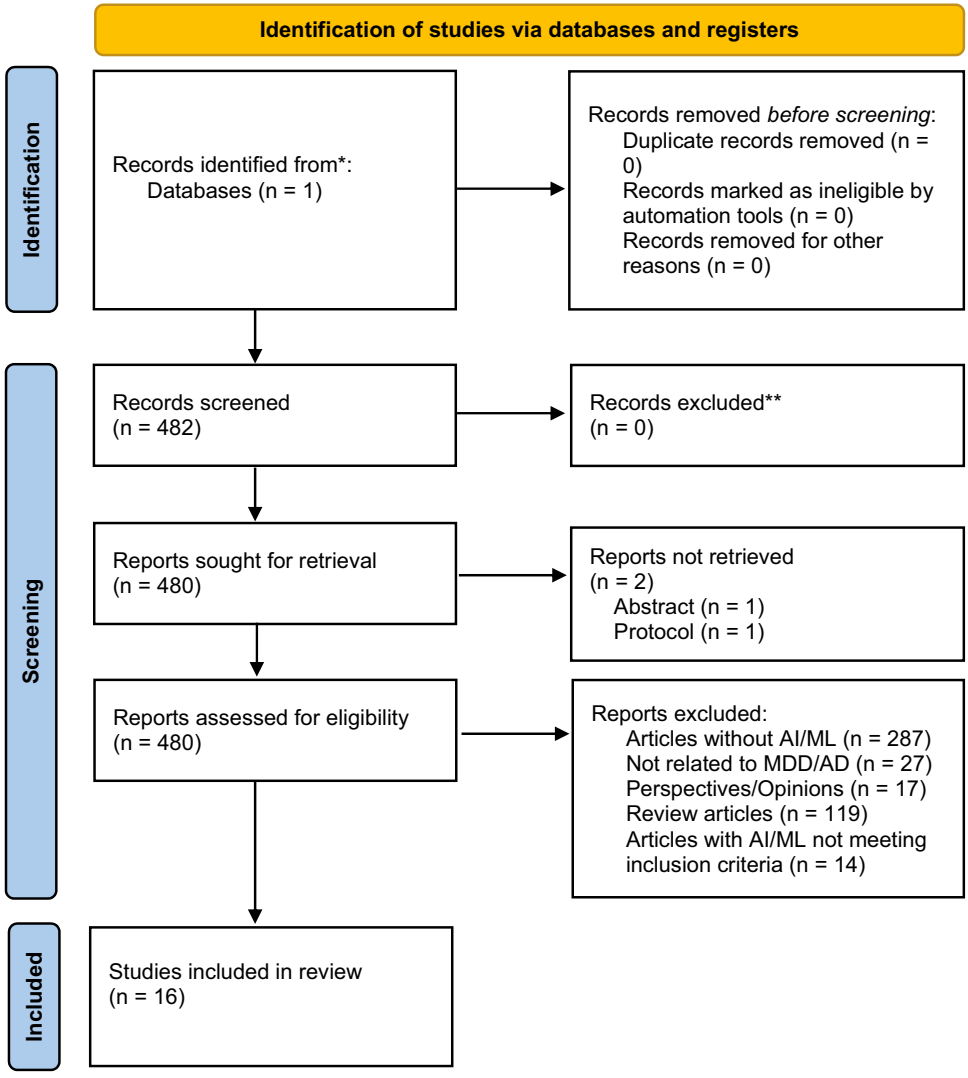

Figure 1. **PRISMA flow diagram**. The flow chart in Figure 1 shows the steps taken to locate individual studies that met criteria for inclusion in this review according to procedures outlined in Guideline of the Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA).

a priori by the investigators or by using an unsupervised learning algorithm (or a combination of the two). The cohorts were typically split into 'training' data used to develop predictive models with cross-validation used for optimizing prediction performance and for minimizing over-fitting of the models to the training data. Toward demonstrating generalizability, some predictive models were further validated using a second or even third independent dataset. The differences in methodological approaches for predicting antidepressant response across studies did not permit a quantitative synthesis of findings. Hence, we review the key methodological characteristics of (Table 2) and the principal findings from each study (Table 3). Outcomes of interest were response, generally defined as a ≥50% reduction (improvement) from baseline in total scores on a depression rating scale, or remission, defined as achieving a certain threshold of improvement in depressive symptom scores during or at the end of follow-up. We will formally define the outcomes of the studies summarized in the next section.

4.3. Models with only genomic predictors

Maciukiewicz and colleagues leveraged support vector machines (SVMs) and classification and regression trees (CARTs) with genomic data to predict response and remission after 8 weeks of treatment with duloxetine in a cohort of 186 patients with MDD [48]. The single nucleotide polymorphisms (SNPs) for this study were selected using logistic regression and Least Absolute Shrinkage and Selection Operator (LASSO) from an initial pool of over 500,000 candidate SNPs. SVM and CART models were trained with nested cross-validation, and there was no validation of trained models in an external dataset. Response was defined as achieving a ≥ 50% reduction in Montgomery Asberg Depression Rating Scale (MADRS) scores and remission was defined as achieving a MADRS score ≤10 at the end of follow-up [49]. The average predictive accuracies of the SVM and CART models (with originally genotyped and imputed variants) were 64% and 57%, respectively, for response and 52% and 51%, respectively, for

Table 2. Methods for machine learning model construction and integration with pharmacogenomic predictors of remission and response with antidepressants in people with major depression.

Lead author/ publication date	Reference number	Data Source(s)[a]	Predictors/ pharmacogenomic feature selection	Machine learning method(s)	Validation method(s)	Interventions/ treatment groups	Depression symptom measures[b]	Predicted outcome
Athreya et al. (2019)	[52]	Clinical trial enrollees: PGRN-AMPS (n = 398) STAR*D (n = 467) ISPC (n = 165)	Pharmacogenomic (six SNPs selected *a priori* in or near *TSPAN5, ERICH3, DEFB1*, and *AHR*) and baseline clinical measures	Supervised (random forest)	Inner 10-fold cross-validation, outer 5-fold cross-validation, with external validation	Citalopram or Escitalopram	QIDS-C HAMD-17	Response Remission
Bao et al. (2021)	[51]	Clinical trial enrollees (n = 83)	Pharmacogenomic (SNPs selected using logistic regression models and random forest)	Supervised (variety of methods were compared)[c]	Inner 10-fold cross-validation, outer 6-fold cross-validation; no external validation	Ketamine (IV)	HAMD	Response
Bi et al. (2021)	[69]	Clinical trial enrollees (n = 610)[d]	Pharmacogenomic (127 markers selected *a priori*), cognitive, neuroendocrine, and personality factors[d]	Supervised (random forest)	Unspecified	SSRIs, SNRIs, NaSSAs, or TCAs[e]	HAMD	Response Remission
Fabbri et al. (2020)	[71]	TRD subjects (GSRD cohort; n = 847 for training, n = 362 for testing), and validation in STAR*D (n = 266) and GENDEP (n = 321)	Pharmacogenomic variants, clinical and demographic factors	GBM	5-fold Cross-validation with 100 repeats; validation in 30% hold-out samples; external validation in STAR*D	SSRI, SNRI	MADRS	Response[f]
Fabbri et al. (2017)	[65]	Pooled European samples (n = 671) with MDD STAR*D subjects with MDD (n = 1,409) STAR*D subjects with TRD (n = 620)	Pharmacogenomic variants in CACNA1C, CACNB2, ANK3, GRM7, TCF4, ITIH3, SYNE1, FKBP5 genes (no subsequent feature selection), clinical and demographic characteristics	Supervised (neural networks, recursive partitioning, learning vector quantization, gradient boosted machine and random forest)	10-fold cross validation with 100 repeats, with external validation	SSRIs or SNRIs	QIDS-C MADRS	In MDD samples: Response, Remission In TRD samples: Non-response
Iniesta et al. (2018)	[66]	GENDEP (n = 430)	Pharmacogenomic (11 [ESC] and 20 [NTP] SNPs selected based on CAT score), sociodemographic 74[ESC], and baseline cli59nical measures [ESC72]	Supervised (LR with elastic net penalty)	5-fold cross-validation with replication in independent (non-overlapping) group of participants	Escitalopram Nortriptyline	HAMD-17	Remission
Joyce et al. (2021)	[74]	Clinical trial enrollees: PGRN-AMPS (n = 264) CO-MED (n = 111)	Pharmacogenomic (six SNPs selected *a priori* in or near *TSPAN5, ERICH3, DEFB1*, and *AHR*), metabolomic, and baseline clinical and sociodemographic measures	Supervised (LR with I2 penalty and XGBoost ensembles)	Nested cross-validation (inner 5-fold cross-validation, outer 3-fold cross-validation) with external validation	Citalopram or Escitalopram	QIDS-C	Response Remission
Kautzky et al. (2015)	[59]	Prospective sample: GSRD (n = 225)	Pharmacogenomic (12 SNPs selected *a priori* in or near *HTR2A, COMT, ST8SIA2, PPP3CC*, and *BDNF*) and clinical measures	Supervised (random forest) and Unsupervised (clustering)	10-fold cross-validation; no external validation	SSRIs, SNRIs, TCA, NARIs, MAOIs, medication combinations, or ECT	HAMD	Response
Lim et al. (2014)	[72]	Naturalistic observation study of MDD subjects with training sample (n = 239, SSRI), testing sample (n = 176, SSRI), and an additional validation sample (n = 234, SNRI)	Pharmacogenomic variants related to serotonin synthesis, serotonin transport, glutamate receptors, and GABA synthesis, clinical, and demographic factors	Logistic regression	External validation in subjects treated with non-SSRI antidepressants	SSRI	HAMD-17	Response

(*Continued*)

Table 2. (Continued).

Lead author/ publication date	Reference number	Data Source(s)[a]	Predictors/ pharmacogenomic feature selection	Machine learning method(s)	Validation method(s)	Interventions/ treatment groups	Depression symptom measures[b]	Predicted outcome
Lin et al. (2020)	[63]	Prospective sample (n = 421)	Pharmacogenomic (10 SNPs in or near *ABCA13, BNIP3, CACNA1E, EXOC4, GRIN2B, LHFPL3, NELL1, NUAK1, PREX1*, and *SLIT3* selected using LR), clinical, and demographic characteristics	Supervised (variety of methods were compared)[g]	10-fold cross-validation; no external validation	SSRIs[h]	HAMD-21	Response Remission
Lin et al. (2018)	[62]	Prospective sample (n = 455)	Pharmacogenomic (10 SNPs in or near *ABCA13, BNIP3, CACNA1E, EXOC4, GRIN2B, LHFPL3, NELL1, NUAK1, PREX1*, and *SLIT3* selected using LR), clinical, and demographic characteristics	Supervised (deep learning using MFNNs)	10-fold cross-validation; no external validation	SSRIs[h]	HAMD-21	Response Remission
Maciukiewicz et al. (2018)	[48]	Clinical trial enrollees (n = 186)	Pharmacogenomic (SNPs selected using logistic and LASSO regression models)	Supervised (SVM and CART)	Nested cross-validation (inner 10-fold cross validation, outer 5-fold cross validation); no external validation	Duloxetine	MADRS	Response Remission
Pei et al. (2020)	[70]	Prospective sample (n = 98 inpatients)	Pharmacogenomic (13 SNPs from an *a priori* list, narrowed using LR) and neuroimaging	Supervised (SVM)	Leave-one-out cross-validation; no external validation	SSRIs or SNRIs[j]	HAMD-6	Response
Shumake et al. (2021)	[68]	Clinical trial enrollees: STAR*D (n = 1,257)	Pharmacogenomic (11 SNPs selected *a priori* and additional SNPs selected using an elastic net)	Supervised (stacked ensemble model)	14-fold cross-validation; no external validation	Citalopram	QIDS-C	Response
Taliaz et al. (2021)	[64]	Clinical trial enrollees: PGRN-AMPS (n = 529) STAR*D (n-1,953)	Pharmacogenomic (selected a priori and then subjected to feature selection using elastic net and LASSO), clinical, and demographic characteristics	Supervised (variety of methods were compared)[i]	5- or 10-fold repeated cross-validations with external validation	Citalopram Sertraline Venlafaxine	QIDS-C	Response
Yin et al. (2015)	[73]	Naturalistic observation study (n = 290)	Pharmacogenomic variants in TPH, SLC6A2, SLC6A3, or DRD2 genes, clinical, neuroendocrine factors, and demographic characteristics	Logistic regression	No external validation	SSRI	HAMD	Remission and Response

Key: CART = classification and regression trees; CAT = Correlation-adjusted T-scores; ECT = electroconvulsive therapy; ESC = escitalopram; IV = intravenous; LASSO = Least Absolute Shrinkage and Selection Operator; LR = logistic regression; MDD = major depressive disorder; MFNN = multilayer feedforward neural networks; NARI = norepinephrine reuptake inhibitor; NaSSA = noradrenergic and specific serotonergic antidepressant; NTP = nortriptyline; SNP = single nucleotide polymorphism; SNRI = serotonin-norepinephrine reuptake inhibitor antidepressant; SSRI = selective serotonin reuptake inhibitor antidepressant; SVMs = support vector machine classification; TCA = tricyclic antidepressant; XGBoost = extreme gradient-boosted decision tree-based ensembles.

[a]Data sources for selected studies included prospective studies such as the Combined Medications to Enhance Outcomes of Antidepressant Therapy (CO-MED) trial, the Genome-based Therapeutic Drugs for Depression (GENDEP) study, the Group for the Study of Resistant Depression (GSRD) study, the Pharmacogenomics Research Network Antidepressant Medical Pharmacogenomic Study (PGRN-AMPS) and the Sequenced Treatment Alternatives to Relieve Depression phase 1 (STAR*D) trial; and pooled clinical trial data from the International SSRI Pharmacogenomics Consortium (ISPC).

[b]Depression symptom measures included the 21-, 17-, or 6-item versions of the Hamilton Depression Rating Scale (HAMD-21, HAMD-17, HAMD-6), the full or unspecified version of the HAMD (HAMD), the Montgomery-Asberg Depression Rating Scale (MADRS), and the clinician-rated version of the Quick Inventory of Depressive Symptomatology (QIDS-C).

[c]The following machine learning algorithms were compared: SVM, decision tree, K-nearest neighbor, random forests, logistic regression with l2 penalty, and logistic regression with elastic net penalty.

[d]Non-pharmacogenomic predictors included measures that assessed selected neurocognitive domains (Wisconsin Card Sorting Test [WCST], Tower of Hanoi, Trails A and B-M, Verbal Fluency Test, and linguistic and operational scales of the Chinese version of the Wechsler Adult Intelligence Scale-Revised), peripheral blood neuroendocrine markers (morning cortisol, adrenocorticotropic hormone [ACTH], thyroid stimulating hormone [TSH], T3, T4, free T3, and free T4), and personality features (as assessed by the Minnesota Multiphasic Personality Inventory [MMPI]) involved in depression.

[e]Patients with MDD were randomized to one of four treatment groups: SSRIs (48%, including citalopram, fluoxetine, paroxetine, or sertraline); SNRIs (32%, including duloxetine or venlafaxine); NaSSA (8%, mirtazapine); or TCAs (12%, including amitriptyline, doxepin, or imipramine).

[f]Response was defined as a MADRS score < 22 and > 50% reduction in MADRS scores from baseline.

[g]The following machine learning algorithms were compared with an ensemble machine learning framework: MFNNs, LR, SVM, C4.5 decision tree, naïve Bayes, and random forests.

[h]Patients with MDD were treated with citalopram, escitalopram, fluoxetine, or paroxetine.

[i]The following machine learning algorithms were compared: SVM, XGBoost, random forest, and Adaptive Boosting (AdaBoost).

[j]Patients with MDD were treated with SSRIs (64%, including escitalopram, fluoxetine, fluvoxamine, paroxetine, or sertraline) or SNRIs (36%, including duloxetine or venlafaxine) at the discretion of the treating clinician.

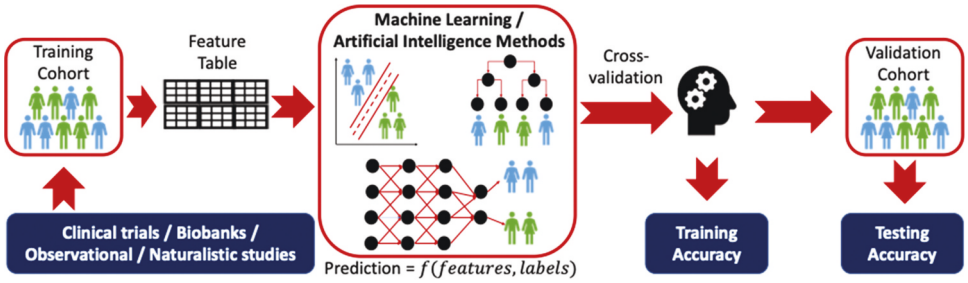

Figure 2. **A general machine learning workflow**. Data from clinical trials or observational studies are converted to feature tables (e.g. individual patients in rows, predictor variables in columns, etc.). Predictive methods are then trained using training data and the training prediction performance metrics are reported. The prediction performance of trained methods is then validated in an independent validation cohort (consisting of either a 'hold out' segment of the original dataset or a separate cohort of patients in another dataset). The prediction performance metrics in the validation cohort are also reported.

remission and not significantly better than chance. Adding baseline depressive symptom scores to the SNP data did not significantly improve the predictive accuracies of the models for response or remission.

Bao et al. [50] compared the performances of six machine learning algorithms that included genomic data for predicting response in 83 patients with MDD who received 2 weeks of treatment with low-dose intravenous (IV) ketamine. All included patients had responded poorly to two or more adequate trials of antidepressants and had active suicidal ideations. The algorithms of interest included SVMs, random forests (RF), k-nearest neighbor (kNN), logistic regression with l2 penalty (LR/l2), decision trees, and logistic regression with elastic net penalty (LR/EN). Genomic SNPs were selected using logistic regression and RFs. All models were trained using nested cross-validation, and there was no validation of prediction models in an external dataset. Response was defined as a ≥50% reduction from baseline in Hamilton Depression Rating Scale (HAMD) scores [51]. Remission was defined as a HAMD score ≤7 at the end of follow-up. The average predictive accuracies for response and remission were generally higher with SVM, kNN, LR/l2, and LR/EN (62%–63%) than with RF and decision trees (56%–57%). The most accurate predictive model was the SVM in fold 6, which had an accuracy of 85% and AUC of 0.86. All of the models tested in this study performed better than a comparator model that used randomized labeled data for responders.

4.4. Models with genomic predictors and clinical/demographic measures

Athreya et al. [52] used unsupervised (clustering) and RFs to predict response and remission in outpatients with MDD who were treated with citalopram or escitalopram for 8 weeks. Psychiatric diagnoses in the testing dataset were confirmed using the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (SCID), 4th Edition. The RFs were trained using baseline depression scores and six SNPs in or near *TSPAN5* (rs10516436), *ERICH3* (rs696692), *DEFB1* (rs5743467, rs2741130, and rs2702877), and *AHR* (rs17137566). Each of these SNPs were previously selected using genome-wide association studies (GWASs) with plasma serotonin and kynurenine concentrations as phenotypes, followed by functional validation in experimental models [53,54]. Data from the Mayo Clinic Pharmacogenomics Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS [55]) was used to train the models on two separate depression rating scales (the 17-item Hamilton Depression Rating Scale [HAMD-17] and the Quick Inventory of Depressive Symptomatology [QIDS]) [51,56]. The random forest models were trained using nested cross-validation. Data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial and the International SSRI Pharmacogenomics Consortium (ISPC) project were used for external validation [57,58]. Across datasets, response was defined as a ≥ 50% reduction in depression scale total scores and remission was defined as a QIDS score ≤5 or HAMD-17 score ≤7 at either 4- or 8 weeks. The study results were stratified by sex. The predictive model trained using PGRN-AMPS data achieved comparable predictive accuracies for response (75%–77% with the HAMD-17 and 66%–69% with the QIDS) and remission (74%–76% with the HAMD-17 and 66%–75% with the QIDS) in the STAR*D and ISPC datasets. The accuracies of the predictive models were higher than the null information rate (NIR), particularly for remission. The NIR, defined as the fraction of the larger class of patients that achieved the outcome of interest (e.g. if 60% of patients were responders, then NIR is 0.60), served as a proxy for chance.

Kautzky and colleagues similarly leveraged clustering and RFs to predict response after 4 weeks of treatment with antidepressants or electroconvulsive therapy (ECT) using 12 SNPs in or near *HTR2A* (rs643627, rs6313), *COMT* (rs4680), *ST8SIA2* (rs8035760, rs3784723), *PPP3CC* (rs7430, rs10108011), and *BDNF* (rs6265, rs11030101, rs11030104, and rs12273363) in 225 depressed participants in the Group for the Study of Resistant Depression (GSRD) cohort [59,60]. MDD diagnoses were confirmed using a modified version of the Mini-International Neuropsychiatric Interview (MINI), version 5.0.0 [61]. SNPs were selected based on literature review. Study drugs included selective serotonin reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), noradrenaline reuptake inhibitors (NARIs), tricyclic and tetracyclic antidepressants, and monoamine reuptake inhibitors (MAOIs). There was no stratification based on antidepressants or intervention types. Response to treatment was defined as achieving a HAMD score ≤17 after one or two adequate trials of

Table 3. Prediction performance of integrated machine learning and pharmacogenomics models for response and remission with antidepressants in people with major depressive disorder[a].

Lead author/ publication date	Reference number	Rating scale and treatment duration	Internal validation				External validation			
			Response		*Remission*		*Response*		*Remission*	
Athreya et al. (2019)[b]	[52]	HAMD-17, 8 wks	ACC	0.86–0.88	ACC	0.83–0.86	ACC	0.75–0.77	ACC	0.74–0.76
			B-ACC	0.86–0.88	B-ACC	0.84–0.87	B-ACC	0.73–0.76	B-ACC	0.74–0.76
			AUC	0.88–0.90	AUC	0.84–0.90	AUC	…	AUC	…
			Sensitivity	0.90[b]	Sensitivity	0.87–0.90	Sensitivity	0.78–0.80	Sensitivity	0.76–0.80
			Specificity	0.82–0.85	Specificity	0.80–0.84	Specificity	0.68–0.71	Specificity	0.71–0.72
			PPV	0.91–0.93	PPV	0.82–0.87	PPV	0.82–0.85	PPV	0.74–0.77
			NPV	0.70–0.79	NPV	0.85–0.87	NPV	0.62–0.63	NPV	0.73[b]
		QIDS, 8 wks	ACC	0.73–0.74	ACC	0.69–0.78	ACC	0.66–0.69	ACC	0.66–0.75
			B-ACC	0.74–0.76	B-ACC	0.72–0.78	B-ACC	0.66–0.70	B-ACC	0.73–0.74
			AUC	0.70–0.85	AUC	0.08–0.75	AUC	…	AUC	…
			Sensitivity	0.70–0.71	Sensitivity	0.60–0.81	Sensitivity	0.67–0.68	Sensitivity	0.59–0.79
			Specificity	0.78–0.80	Specificity	0.75–0.83	Specificity	0.63–0.72	Specificity	0.69–0.72
			PPV	0.85–0.86	PPV	0.84[b]	PPV	0.78–0.82	PPV	0.67–0.75
			NPV	0.57–0.62	NPV	0.59–0.69	NPV	0.51–0.52	NPV	0.63–0.72
Bao et al. (2021)[c]	[51]	HAMD, 2 weeks	ACC	0.62	…	…	…	…	…	…
			AUC	0.59						
Bi et al. (2021)[d]	[69]	HAMD-17, 6 weeks	AUC	0.75–0.77	…	…	…	…	…	…
Fabbri et al. (2020)	[71]	MADRS[e]	AUC	0.65–0.75	…	…	AUC[f]	0.60–0.62	…	…
							AUC[g]	0.55–0.72		
Fabri et al. (2017)	[65]	QIDS, HDRS, MADRS, 4 weeks	ACC	0.73	…	…	…	…	…	…
			Sensitivity	0.83						
			Specificity	0.56						
Iniesta et al. (2018)[h]	[66]	HAMD-17, 12 weeks	…	…	AUC	0.77–0.83	…	…	AUC	0.77 [h]
					B-ACC	0.74–0.77			B-ACC	0.70–0.78
					Sensitivity	0.69–0.70			Sensitivity	0.69–0.68
					Specificity	0.77–0.83			Specificity	0.71–0.87
Joyce et al. (2021)[i]	[74]	QIDS, 8 weeks	ACC	0.67	…	…	ACC	0.76	…	…
			AUC	0.68			B-ACC	0.79		
							AUC	0.83		
							Sensitivity	0.69		
							Specificity	0.88		
Kautzky et al. (2015)	[59]	HAMD, 4 weeks	Average MDA	5.6	…	…	OR	4.22	…	…
							(95% CI)[j]	(1.43–12.49)		
Lim et al. (2014)	[72]	HAMD-17. 6 weeks	ACC	.87[k]	…	…	ACC	0.85	…	…
Lin et al. (2020)[l]	[63]	HAMD-21, 8 weeks	AUC	0.82–0.83	AUC	0.81 [l]	…	…	…	…
			B-ACC	0.73–0.74	B-ACC	0.72 [l]				
			Sensitivity	0.75–0.77	Sensitivity	0.78 [l]				
			Specificity	0.71 [l]	Specificity	0.65–0.66				
Lin et al. (2018)	[62]	HAMD-21, 8 weeks	AUC	0.82	AUC	0.81	…	…	…	…
			B-ACC	0.72	B-ACC	0.72				
			Sensitivity	0.75	Sensitivity	0.77				
			Specificity	0.69	Specificity	0.66				
Maciukiewicz et al. (2018)[m]	[48]	MADRS, 8 weeks	ACC	0.64	…	…	ACC	0.52	…	…
			B-ACC	0.47			B-ACC	0.52		
			Sensitivity	0.87			Sensitivity	0.58		
			Specificity	0.07			Specificity	0.46		
Pei et al. (2020)	[70]	HAMD-6, 2 weeks	ACC	0.86	…	…	…	…	…	…
			B-ACC	0.86						
			Sensitivity	0.87						
			Specificity	0.84						

(*Continued*)

Table 3. (Continued).

Lead author/ publication date	Reference number	Rating scale and treatment duration	Internal validation				External validation			
			Response		*Remission*		*Response*		*Remission*	
Shumake et al. (2021)[n]	[68]	QIDS, 14 weeks	ACC	0.63	…	…	…	…	…	…
			AUC	0.66						
			B-ACC	0.64						
			Sensitivity	0.64						
			Specificity	0.61						
Taliaz et al. (2021)	[64]	QIDS, 14 weeks	B-ACC	0.72	…	…	B-ACC	0.61	…	…
Yin et al. (2015)	[73]	HAMD-17, 6 weeks	VAR	0.75	VAR	0.66	…	…	…	…

Key: ACC = accuracy; AUC = area under the receiver operating curve; B-ACC = balanced accuracy (calculated as [sensitivity + specificity]/2); NPV = negative predictive value; HAMD = Hamilton Depression Rating Scale (21-, 17-, and 6-item versions, and version unspecified); MDA = Mean Decrease in Accuracy; OR = odds ratio; PPV = positive predictive value; QIDS = Quick Inventory of Depressive Symptomatology; SNP = single nucleotide polymorphism; VAR = percentage of variation in outcome (response or remission) predicted by the model.

[a]Performance measure results are rounded to two significant figures to the right of the decimal point.

[b]Ranges of performance measures are reported for Athreya et al. (2019) because the prediction performances of random forests were reported as sex-specific ACCs, AUCs, B-ACCs, sensitivities, specificities, NPVs, and PPVs. In the training dataset, sensitivity using the HAMD-17 and PPV using the QIDS were the same for women and men. In the testing dataset, NPV using the HAMD-17 was the same for women and men.

[c]Results are reported for SVM model as average accuracy and AUC.

[d]Ranges of AUCs are displayed because the prediction models were run separately for SSRI response and SNRI response.

[e]The gradient boosted machine model (GBM) was leveraged to predict lack of response to two or more adequate trials of various antidepressants.

[f]Results shown here are for model performance in the Genome-based Therapeutic Drugs for Depression study (GENDEP)-derived validation sample

[g]Results shown here are for model performance in the Sequenced Treatment Alternatives to Relieve Depression study (STAR*D)-derived validation sample.

[h]Ranges of performance measures are displayed for the report by Iniesta et al. (2018) because the prediction performance measures for separate drug-specific elastic net logistic regression models are reported for patients who were treated with escitalopram and patients who were treated with nortriptyline. Cross-drug performance measures are not reported in this table. In the testing dataset, the AUC using the HAMD-17 was the same for escitalopram- and nortriptyline-treated patients.

[i]Results are shown for extreme gradient-boosted decision tree-based ensembles (XGBoost) using a 'multiomics model' comprised of genomic, metabolomic, and clinical predictors.

[j]Patients with GG–TT–GG for rs6265–rs6313–rs7430 and no melancholia had 4-fold higher odds of being responders compared to the rest of the patients in the sample.

[k]Results were reported for the 60% of the total sample for which the LR model with I2 penalty gave an outcome prediction., these results could not be replicated in a cross-validation dataset consisting of depressed patients who were treated with non-SSRI antidepressants.

[l]Results are shown for multilayer feedforward neural network models with 2 hidden layers (response) and with 3 hidden layers (remission). Ranges of performance measures are reported because prediction performance measures were compared for boosting ensemble models with and without feature selection. For prediction of response, specificity was the same for boosting ensemble models with and without feature selection after rounding. For prediction of remission, AUC, B-ACC, and sensitivity were the same for boosting ensemble models with and without feature selection after rounding.

[m]Results are shown for support vector machine models consisting or both imputed and originally-genotyped variants.

[n]Results are shown for prediction performance of ensemble learning using a combination of clinical predictors and SNPs that were selected *a priori*.

antidepressants. The RF model was trained using 10-fold cross-validation, and the trained models were not validated in an external dataset. A four-factor RF model incorporating three SNPs (rs6265, rs6313, and rs7430) and melancholic depressive subtype was associated with a 4-fold higher chance of positive treatment response compared with other patients (OR 4.22, 95% CI 1.43–12.49).

Lin et al. [62] and coauthors tested the performance of supervised deep learning using multi-layer feedforward neural networks (MFNNs) for predicting response and remission after 8 weeks of treatment with SSRIs in a prospective cohort of outpatients with an investigator-confirmed diagnosis of MDD. Predictor variables for these models included clinical and demographic characteristics and SNPs in or near *ABCA13* (rs4917029), *BNIP3* (rs9419139), *CACNA1E* (rs704329), *EXOC4* (rs6978272), *GRIN2B* (rs7954376), *LHFPL3* (rs4352778), *NELL1* (rs2139423), *NUAK1* (rs2956406), *PREX1* (rs4810894), and *SLIT3* (rs139863958). Response was defined as ≥50% reduction from baseline in 21-item HAMD (HAMD-21) scores, and remission was defined as a HAMD-21 score ≤7 at week 8. MFNNs containing 1–3 hidden layers and MFNNs with logistic regression were trained using 10-fold cross-validation, and the trained models were not validated in an external dataset. MFNNs containing two hidden layers achieved the highest AUC (0.82), sensitivity (0.75), and specificity (0.69), although all models performed almost equally well.

These investigators subsequently compared a variety of supervised learning models for predicting response (≥50% reduction in HAMD-21 score) and remission (HAMD-21 score ≤7) after 8 weeks of treatment with SSRIs using a similar cohort of depressed patients [63]. The predictive models included boosted ensemble machine learning, neural networks (MFNNs), logistic regression, SVMs, decision trees, RFs, and naïve Bayes models. The clinical, demographic, and genomic predictors used in this study were similar to the predictors used in their previous study [62]. Of the models tested, the boosting ensemble algorithm with feature selection achieved the highest AUC (0.81–0.83), sensitivity (0.77–0.78), and specificity (0.66–0.71) for predicting both response and remission, although all models generally performed well (minimum AUCs for predicting response and remission were 0.68 and 0.63, respectively). None of these models were further validated using an external dataset.

Taliaz et al. [64] also compared several supervised learning models for predicting response to 14 weeks of treatment with citalopram, sertraline, or venlafaxine using clinical and demographic variables and genomic predictors. Genomic predictors were initially selected *a priori* based on literature review and were then used as predictors in elastic net logistic regression with Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection. The training sample included 1,953 patients with MDD who participated in the STAR*D trial. An external validation sample included 529 patients from the PGRN-AMPS trial. Since the PGRN-AMPS trial included only citalopram as a study drug, only the citalopram algorithm from STAR*D was validated using PGRN-AMPS data. Supervised learning methods included SVMs, extreme gradient-boosted decision tree-based ensembles (XGBoost), RFs, and Adaptive Boosting (AdaBoost). Of these, SVMs with a linear kernel achieved the best prediction performance in the STAR*D training dataset (average balanced accuracy 73% across medications) and in the PGRN-AMPS validation dataset (average balanced accuracy 72%). The balanced accuracy for the algorithm's citalopram model was similar in the STAR*D (60.5%) and PGRN-AMPS datasets (61.3%).

Fabbri and colleagues investigated the accuracies of machine learning models that combined genomic, clinical, and sociodemographic factors for predicting response, remission, and treatment-resistance in patients with MDD after 4 weeks of treatment with SSRIs or SNRIs [65]. The study datasets included patients with MDD from pooled European samples and a separate dataset consisting of STAR*D participants for external validation. Genomic predictors included 44 SNPs in or near *CACNA1C, CACNB2, ANK3, GRM7, TCF4, ITIH3, SYNE1*, and *FKBP5* that were chosen by the investigators. Machine learning models included neural networks, recursive partitioning, learning vector quantization, gradient boosted machines (GBMs), and RFs. When combined with clinical and demographic characteristics, the best-performing candidate genes (*ANK3, CADNB2, FKBP5*, and *CACNA1C*) for predicting response (≥50% reduction in HAMD-21 or MADRS scores at weeks 4 or 6), remission (score ≤7 on the HAMD-21 or <10 on the MADRS), or treatment resistance (poor response to at least two consecutive antidepressant trials of adequate design) in the European datasets were tested for associations with response and remission in treatment-resistant STAR*D patients. Neural networks and GBMs had the highest predictive accuracies among the models tested in STAR*D (mean Accuracy 73%, Sensitivity 0.83, Specificity 0.56), although predictive performances did not differ greatly across all machine learning algorithms.

Iniesta et al. [66] used supervised learning (LR/EN) with genomic predictors, sociodemographic variables, baseline depressive symptom measures, and other clinical variables to predict remission after 12 weeks of treatment with escitalopram or nortriptyline. Data for this study were from 430 patients with MDD who participated in the Genome-based Therapeutic Drugs for Depression (GENDEP) study [67]. Variable selection was conducted using 5-fold cross-validation in the training data, with further validation in a non-overlapping group of GENDEP participants. Correlation-Adjusted T (CAT) scores were used to select predictors, including the final set of SNPs (11 for escitalopram and 20 for nortriptyline) from an initial pool of over 500,000. Remission was defined as a HAMD-17 score of ≤7 at the last observation after at least 4 weeks. For escitalopram, the LR/EN models achieved high predictive performances for remission in both the training (AUC 0.80, Sensitivity 0.71, Specificity 0.77) and testing datasets (AUC 0.77, Sensitivity 0.69, Specificity 0.71). For the nortriptyline model, there were similarly high predictive performances for remission in the training (AUC 0.83, Sensitivity 0.70, Specificity 0.83) and testing datasets (AUC 0.77, Sensitivity 0.68, Specificity 0.87). However, the model for escitalopram achieved an AUC, sensitivity, and specificity of 0.57, 0.46, and 0.67, respectively, when applied to nortriptyline-treated patients in cross-drug sensitivity analyses. The performance of the nortriptyline model was similarly poor for predicting remission in escitalopram-treated patients.

Shumake et al. [68] used a stacked ensemble machine learning model to predict response to citalopram using data from 1,257 STAR*D participants. MDD diagnoses were verified using a symptom checklist. All subjects were followed for up to 14 weeks. Genomic predictors included 11 SNPs (rs1392611, rs10812099, rs1891943, rs151139256, rs11002001, rs62182022, rs28373080, rs7757702, rs76557116, rs9557363, rs2704022) that were selected *a priori* based on literature review. Additional genomic predictors were selected using LR/EN from an initial pool of over 350,000 candidate SNPs. Clinical predictors included sociodemographic and baseline depression symptom measures. Adequate treatment response was defined as a QIDS score ≤5. Performance measures were calculated for models that included: (a) only clinical and sociodemographic predictors; (b) *a priori* SNPs in addition to clinical and sociodemographic predictors; and (c) elastic net SNPs combined with clinical and sociodemographic predictors. The best-performing model combined *a priori* SNPs with clinical and sociodemographic predictors (Accuracy 63%, AUC 0.66); however, the remaining models performed nearly as well (Accuracy 61%–62%, AUC 0.66). A model that included only genetic variants (without clinical and sociodemographic predictors) did not outperform chance predictions. None of the models were further validated in external datasets.

Bi and coauthors tested an RF algorithm for predicting response and remission in a dataset that consisted of 610 patients with MDD who received 6 weeks of treatment with SSRIs, SNRIs, and other antidepressants [69]. Machine learning models focused on predicting treatment outcomes with SSRIs and SNRIs. Genomic predictors included 127 markers that were selected *a priori* by the investigators based on their possible involvement in the pharmacodynamic activities of antidepressants; however, the final prediction models included only three SNPs (rs13353402, rs17289304, and rs32897). Additional predictors included anxiety symptoms, cognitive factors, and peripheral blood neuroendocrine markers, although neuroendocrine markers included only in the final SNRI models and not the SSRI models. Response was defined as >50% reduction in HAMD-17 scores and remission was defined as a HAMD-17 score <8 during follow-up. The AUCs of the SSRI and SNRI models were 0.77 and 0.75, respectively. Internal validation methods for these models were unspecified, and there was no further validation in a separate dataset.

Pei et al. [70] used an SVM algorithm to predict early response to treatment with SSRIs or SNRIs in a prospective sample of 98 hospitalized patients with MDD. Predictors included SNPs chosen *a priori* by the investigators based on literature review (narrowed to 13 SNPs using logistic regression), sociodemographic measures, and neuroimaging (resting functional connectivity) data. Early response was defined as a ≥ 50% reduction in 6-item HAMD (HAMD-6) total scores from baseline to 2 weeks. The SVM algorithm was trained using leave-one-out cross-validation, and there was no validation of the trained model using an external dataset. The investigators compared the predictive performances of SVM models that included: (a) only neuroimaging + sociodemographic data; (b) only genomic + sociodemographic data, and (c) combined neuroimaging data + genomic data + sociodemographic data. The predictive performance of the model that included neuroimaging + genomic + sociodemographic data (Accuracy 86%, Sensitivity 87%, Specificity 84%) was slightly higher than models with only neuroimaging + sociodemographic data (Accuracy 81%, Sensitivity 78%, Specificity 84%) and only genomic + sociodemographic data (Accuracy 73%, Sensitivity 74%, Specificity 71%).

Fabbri et al. [71] leveraged a GBM algorithm to predict lack of response to two or more adequate trials of various antidepressants in a cohort of patients with MDD who participated in the GSRD project. In the training dataset, genomic, clinical, and sociodemographic predictors were trained using 5-fold cross-validation with 20 repeats (for genomic data) and 100 repeats (for pathway-based scores). Predictors were selected for the GBM algorithm inclusion based on a local false discovery rate <0.8 in at least 50% of the repeats. The GBM algorithm was then trained using 5-fold cross-validation, with further validation in a hold-out sample and external STAR*D and GENDEP datasets. Response was defined as a MADRS score <22 and >50% reduction in MADRS scores from the baseline. Models that combined genomic predictors with clinical and demographic predictors, with AUC 0.65–0.75 in the testing sample, and AUC 0.60–0.62 and 0.55–0.72, respectively, in GENDEP and STAR*D datasets.

Lim et al. [72] trained and tested an LR model with l2 penalty that used a combination of demographic variables and genomic variants related to serotonin synthesis, serotonin transport, glutamate receptor function, and GABA synthesis to predict response to SSRIs in a naturalistic cohort of patients with MDD after 6 weeks of treatment. MDD diagnoses were all confirmed using a structured clinical interview. Response was defined as >50% reduction in HAMD-17 scores from baseline. The 155 SNPs in this study were selected *a priori* based on literature review. The trained LR model was validated externally in a separate group of 176 patients who were treated with SSRIs. For the 60% of the total cases where the model gave a prediction, the predictive accuracy for response was 87%, compared with the posterior probability of 66% for response. However, these results were not validated in a dataset consisting of depressed patients who were treated with non-SSRI antidepressants.

Yin and colleagues developed an LR model with l2 penalization for predicting remission and response to SSRIs in a naturalistic cohort of 290 depressed patients who were treated for 6 weeks [73]. Model predictors included clinical and demographic factors, neuroendocrine markers, and 19 SNPs selected by the investigators in or near *TH, DRD2, DRD4, SLC6A2*, and *SLC6A3* genes. Response was defined as >50% reduction in HAMD-17 scores from baseline; remission was defined as a HAMD-17 score ≤7 at week 6. The model combining all four groups of factors predicted 75% of the variation in response to SSRIs and 66% of variation in remission. There was no external validation.

4.5. Models with multi-omics (genomic and other -omics) predictors and clinical/demographic predictors

To our knowledge, only one study combined genomic, other -omics, and clinical/demographic predictors of antidepressant effects within a machine learning framework. Joyce et al. [74]

leveraged supervised machine learning (LR/I2 and XGBoost ensembles) with 5-fold cross validation-based training to predict response and remission in 298 citalopram- and escitalopram-treated patients with MDD who participated in the PGRN-AMPS trial and the Combining Medications to Enhance Outcomes of Antidepressant Therapy (CO-MED) study [75]. The validation cohort consisted of a separate subgroup of 77 depressed CO-MED patients who were treated with antidepressant combinations. Genomic predictors included six functionally validated SNPs located in or near *TSPAN5, ERICH3, DEFB1*, and *AHR*. Additional predictors included baseline clinical and sociodemographic measures as well as 153 targeted metabolites that met quality-control criteria in both the PGRN-AMPS and CO-MED datasets [76–78]. In the training dataset, XGBoost and LR/I2 models achieved accuracies of 67% and 65%, respectively, and AUCs of 0.68 and 0.72, respectively, for predicting response after 8 weeks of treatment with citalopram or escitalopram. In the validation dataset, the models yielded accuracies of 76% and 75%, respectively, and AUCs of 0.83 and 0.86, respectively, for predicting response to a combination of antidepressant treatments. The performances of both the XGBoost and LR/I2 algorithms were higher than the NIR.

5. Expert opinion

5.1. Current state and future needs

Artificial intelligence/machine learning (AI/ML) approaches using clinical and biological data collected over the past two decades are helping the field to take important steps toward the goal of accurately predicting outcomes of treatment with antidepressants in depressed patients [79]. For clinicians, the accurate prediction of antidepressant treatment outcomes at the individual patient level is challenging given the complex interactions of contributing genetic, non-genetic, psychological, and/or environmental factors [24,25,80]. This reality presents both opportunities and challenges for clinicians and researchers. At no point in history has there been available such powerful computing infrastructures and access to such vast repositories of clinical, social, and biological data [39]. As a result, AI/ML can be used to analyze high-dimensional genomic and non-genomic predictors and clinical variables simultaneously in hopes of achieving quantitative, rule-based decision systems with sufficient validity for clinical use [19,81]. Such technological advances can augment clinicians' contextual assessment of symptom severity of the disease and individualize treatment for the patient. This represents an important step toward biologically driven individualized treatment decision-making in psychiatry.

In summary, this review highlights the promise of combining pharmacogenomics data with statistical and AI/ML approaches for predicting short-term treatment outcomes with antidepressants in patients with MDD. The methodological features and results of the reviewed studies varied widely but may be summarized along three general lines: types of features selected for prediction, types of AI/ML approaches used, and the approaches to feature selection. In terms of the types of features selected for prediction, AI/ML algorithms that incorporated only pharmacogenomic biomarkers to predict therapeutic outcomes with antidepressants yielded levels of predictive performance that are similar to AI/ML algorithms that use only clinical/demographic variables as predictors, generally falling in the AUC range of 0.54–0.67 [34–37]. On the other hand, the performance of AI/ML algorithms appeared to improve substantially when a combination of clinical, demographic, and pharmacogenomic variables was used, generally exceeding an AUC threshold of 0.70 despite varying approaches to feature selection and the types of machine learning algorithms used. As of this writing, it is unclear if an integrative pharmacogenomics approach (i.e. combining pharmacogenomics with other -omics data such as epigenomics, transcriptomics, proteomics, and metabolomics with clinical/demographic information) deployed within an AI/ML framework leads to even better response prediction – a question that will be addressed in future studies. Regarding the relative performances of the different AI/ML algorithms in the reviewed studies, relatively few reports provided a direct comparison of approaches within the same cohort. Among studies that provided such comparisons, the performance measures were similar across AI/ML approaches and no single machine learning approach appeared to be clearly superior to the others. The predictive accuracies of the machine learning models also did not vary substantially according to the methods by which pharmacogenomic and other predictive features were selected. Given important differences in the mathematical underpinnings of the different supervised learning and feature selection methods, more comparative studies are needed.

Though promising, the use of AI/ML for the prediction of therapeutic response to antidepressants is still in its relative infancy [82], with or without the use of genomic markers, and many caveats exist that can impact the validity and clinical utility of model predictions. Although very high predictive accuracies can be achieved using machine learning approaches, the performances of the algorithms are highly dependent on clinical context [33,83]. The 'clinical context' of the data is defined by many 'seen' (the actual input data) and 'unseen' factors (unmeasured factors that are not specifically accounted for in the datasets, e.g. active psychosocial factors in patients). Although both types of contextual elements determine the accuracy of machine learning algorithms [83], the 'unseen' elements cannot be 'learned' by the algorithms, which may influence the treatment outcomes more than genetic factors alone. Moreover, the 'seen' data used in machine learning algorithms often undergo multiple transformations, leading to unpredictable behaviors that can be difficult to detect and interpret [84]. Under these conditions, the predictions from machine learning algorithms may become biased or uninterpretable [85], highlighting the importance of both detailed transparency when reporting methods and results and rigorous validation of algorithm performance with replication in independent samples or datasets, as discussed further below.

There are also inherent limitations associated with the use of genomic data for generating valid machine learning models for antidepressant response prediction. Individual genetic loci discovered thus far can explain only a small proportion of the

heritability or variation in antidepressant responses [86]. Therefore, even with powerful machine learning algorithms and high-dimensional genomic data, reliable and valid response prediction at a level considered sufficient for clinical decision-making cannot be achieved using genomic SNPs alone [59]. Using genomics and 'other -omics' together (integrative -omics) to predict the outcomes of antidepressant treatment is one potential solution [87,88]. However, very few groups have taken this approach [74]; therefore, it is not yet clear if an integrative -omics approach results in better prediction than the more common approach of combining clinical measures with selected genomic data. A promising alternative approach is the leveraging of targeted functional genomics, wherein the SNPs used in predictive models are identified using GWAS based on clinically relevant phenotypes followed by functional validation in experimental models. A targeted functional approach was applied in one study that used SNPs in the *DEFB1, AHR, TSPAN5*, and *ERICH3* genes as predictors [26]. These genes were identified using GWAS for serotonin or kynurenine biosynthesis [53,54] – both of which are important factors in determining MDD disease risk and response to antidepressants [89]. These genes were functionally validated in experimental models which showed that knocking down the expression of *TSPAN5* and *ERICH3* in neuronally derived cell lines decreased serotonin concentrations [53] and that the *DEFB1* gene encodes a protein that can inhibit inflammation and kynurenine synthesis [54].

There are other pragmatic considerations and challenges for future research of algorithm-based prediction of antidepressant treatment outcomes. First, as alluded to earlier, future studies are needed that focus on integrating symptom-based factors with wider arrays of predictive biomarkers to facilitate an integrative-omics machine learning approach to predicting clinical outcomes in depressed patients. This will require large-scale efforts at collecting data of sufficient quality to enable the systematic investigation of genomic and other -omic predictors of response to antidepressants. Fortunately, the existing array of high-quality data collected in research environments will be supplemented by similar types of data collected from routine care environments that may be similarly well suited for this purpose [90–92]. However, to date, a unifying framework for conducting future studies of machine learning algorithms for treatment response prediction in depressed patients is lacking, and the field has not yet achieved meaningful consensus on the best means of consolidating existing samples and datasets and standardizing approaches to future data collection efforts.

Second, the clinical datasets used to train and validate integrated machine learning-genomic models reviewed here included predominantly depressed individuals of European or East Asian ancestry. For the potential of genomics and machine learning-informed response prediction to be shared equitably, greater inclusion of under-represented racial and ethnic groups in genomic and other -omics studies is needed [93].

Third, the studies reviewed herein focused on the prediction of response and remission. Both are reasonably validated dichotomous antidepressant response phenotypes. However, in addition to these phenotypes, future studies should include the prediction of other response phenotypes that may be more meaningful to clinicians, patients, and their caregivers, such as recovery and sustained recovery [12,94].

Fourth, all of the reviewed studies focused on the use of AI/ML and genomics to predict outcomes of antidepressant treatment in non-elderly adults. As of this writing, we are unaware of any such studies focused on the prediction of antidepressant effects in depressed children/adolescents or geriatric adults using AI/ML and genomics.

And finally, none of the machine learning models reviewed here were developed for the purposes of antidepressant selection, given a set of baseline factors. Instead, the machine learning models reviewed in this report were developed to predict outcomes of treatment with antidepressants that were already initiated. Future studies will be needed to address these important, pragmatic knowledge gaps.

5.2. Translation to practice: beyond validation

The limitations described above and threats to external validity posed by model overfitting [95] highlight the importance of both validating and replicating the performances of ML prediction models, especially for complex phenotypes like antidepressant response. Even though a 'hold-out' segment of the validation sample may be previously 'unseen' by the ML algorithm, it is still a randomly selected subset of the same dataset from where training data are drawn, thus limiting its rigor as a validation approach. Validation in a separate, external dataset is a more rigorous test of algorithm performance, as it represents a truer test of external validity. Independent replication is a gold standard for validating genetic biomarkers for psychiatric disease risk and response phenotypes, and we suggest that a similar standard for validation may apply analogously to ML algorithms for antidepressant response prediction. However, to date, only a small number of machine learning studies on the prediction of therapeutic outcomes of antidepressant treatment have used external datasets for algorithm validation [52,64,65,71,72,74].

As a related matter, no studies, to our knowledge, have subjected machine learning models developed retrospectively to subsequent prospective validation. Like retrospective validation using a dataset that exists independently of the training data, prospective validation has the advantage of testing the algorithm in a new and independent cohort of depressed patients. However, prospective validation adds an important level of rigor since machine learning-based predictive algorithms would be used prospectively in actual practice, where treatment outcomes are not yet known and where exclusion criteria are less-stringent than those of clinical trials. Moreover, retrospective validation only focuses on replication predictive accuracy, whereas prospective validation enables the examination of additional outcomes, such as clinical utility (from the patient and practitioner perspective) and ease-of-implementation (from the health-care systems perspective). Furthermore, the clinical utility of predictive models will depend on both their absolute performance (i.e. the accuracy of predictions) and their relative performance (the accuracy of predictions relative to an appropriate comparator condition) [96]. Hence, prospective validation enables comparisons

between model performance and a clinician's best estimate as to the eventual outcome of interest and/or how well a predictive algorithm compliments clinical guesswork. In our view, these are more clinically meaningful standards than random chance or the NIR since a clinician's guess incorporates important predictive information in-and-of-itself and because machine learning-based tools compliment–but do not replace–clinical judgment, as discussed below.

5.3. Translation to practice: the need for novel architectures

A common theme across all studies was the use of next-generation sequencing data from research studies, as opposed to using clinical laboratory generated gene panel data for training/testing AI/ML methods. Looking to the future, the adoption of validated machine learning approaches using genomics for predicting antidepressant response will likely require both clinical inputs (e.g. basic demographics, individual-item scores on depression rating scale, etc.) and laboratory-based inputs (e.g. SNP panel results). Therefore, novel architectures will need to be considered to run the machine learning algorithms in clinical care environments. There are at least three possible architectures suitable for this purpose: (1) a laboratory-based workflow for ordering and billing (Figure 3(a)); (2) a Clinical Decision Support electronic health record (EHR) application designed to interface directly with the ordering clinician's EHR (Figure 3(b)); and (3) a custom platform or application that clinicians interact with directly, via a provider-facing application programming interface (API) (Figure 3(C)). Each of these architectures presents its own unique challenges to implementation, billing for service, and ongoing use by providers. Important factors to consider when selecting a model will include ordering clinicians' requirements, regulatory concerns, timelines, information technology (IT) build complexity, the desired market, and scalability to future machine learning approaches using genomics for predicting other treatment response phenotypes for other disease groups.

With these factors in mind, each architecture possesses its own pros and cons. Option 1 follows traditional ordering and resulting mechanisms, which may appeal to clinicians. The relative complexity of implementation would be simple and, on that basis, would be expected to require a shorter timeline to operationalize and would allow for billing in a traditional fee-for-service manner. However, Option 1 (Figure 3A) requires burdensome data entry by the ordering provider, their staff, or patients and would still require a custom interface build to pass data between the laboratory and the algorithm itself. Option 2 (Figure 3B) offers the most seamless workflow integration for clinicians, eliminating the need for manual and double entry of assessment scores. This option would require an extensive IT build of the platform and a high level of engagement with new external organizations to verify that external EHR data would be available in the format required by the algorithm. Option 3 (Figure 3C) would allow for the addition of future AI prediction algorithms for other medications and disease states. However, Option 3 would require a relatively complex IT build that would allow for the direct interface with external providers, would carry a high level of regulatory concerns, and may create access/utilization concerns with providers unwilling to navigate to a platform external to the EHR. Overall, healthcare organizations that are planning to implement a new technology will need to consider their resource availability and weigh the factors indicated in selecting the right-system architecture.

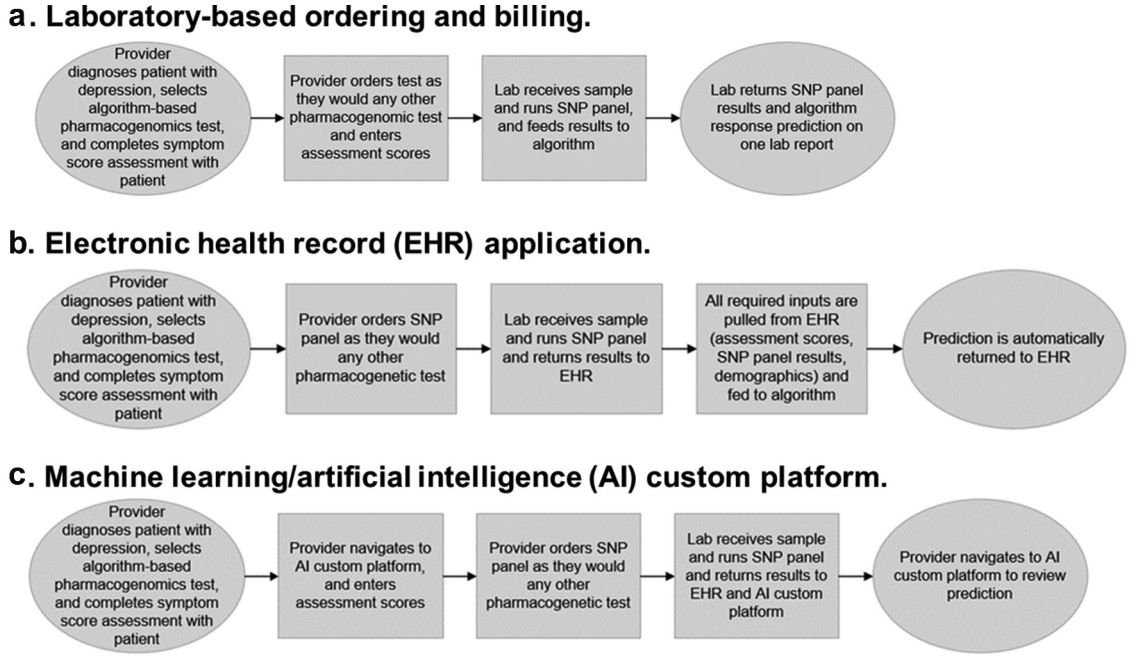

Figure 3. **Novel architectures**. The diagram in Figure 3 illustrates three different approaches to building infrastructures that can facilitate the use of pharmacogenomic and clinical inputs to translate machine learning algorithms for antidepressant response prediction to clinical practice.

5.4. Toward machine learning 'augmented' decision-making

The disruptive potential of genomic testing was recognized well before the launch of the Human Genome Project [97]. As a result, innovations in genomic medicine have often included embedded research studies assessing the psychosocial impact of genomic results [98]. These studies have generated important insights into how best to limit the potential burdens of genetic knowledge, improve patient understanding of genomic results, and integrate genomic results into patient care more effectively. Nonetheless, few studies have examined the impact of preemptive genomic testing, including clinician views of pharmacogenomics [99–102]. A significant limitation of these studies is that they have tended to focus on a highly abstract concepts of pharmacogenomic testing that is disconnected to a specific test, test result, or personal experience with pharmacogenomic testing. As pharmacogenomic testing is increasingly integrated into patient care activities, there remains a significant need to understand clinicians' perceptions of the value of machine learning/AI-enhanced pharmacogenomic prescribing information. Clinicians who treat depressed patients will have multiple concerns regarding the use of genomics data, including its integration with machine learning for guiding treatment decisions. On the genomics side, this will include concerns about the accuracy and relevance of genomic information to specific clinical decisions in caring for depressed patients [103–105]. On the AI/ML side, additional concerns may also include perceptions of liability, limited transparency, or even concerns about reaching erroneous decisions [106,107], given the complexity of machine learning-informed clinical decision support and clinicians' perceptions of professional duties when elements of their practice are guided by sophisticated predictive models that are based on mathematical algorithms that may be largely inscrutable. Patients, in turn, may also question exactly who or what is driving decisions about their treatment. Given the limitations and constraints outlined in this review, we conclude that even the most sophisticated of machine learning algorithms incorporating the best possible sets of predictors will not–and should not–replace the judgment of clinicians. Instead, validated machine learning algorithms will be best viewed as tools that can augment clinical judgment when it comes to predicting the outcomes of treatment in depressed patients.

Finally, the systematic development and implementation of predictive models for drug response in clinical practice is a multidisciplinary effort. In our view, the development of predictive models with sufficient validity for clinical use requires close collaborations between computer scientists, informaticians, biostatisticians, genomics experts, and clinicians. The implementation of validated predictive models will warrant similar collaborations further downstream between these same disciplines, health system engineers, and experts in laboratory medicine. Integrating prediction models at point of care will continue to involve the complex operation of homogenizing data formats from laboratory tests/sequencing panels, creating intuitive order sets and easily interpretable reports within the electronic health records, and most importantly, training the health-care provider workforce on the basics of the methods and the proper interpretations of model outputs. Therefore, understanding the needs and preferences of the clinician end-users prior to the design, implementation, and adoption of technologies is likely to improve the trust and uptake of predictive tools in busy clinical practices. We thus assert that the successful development and adoption of predictive methods utilizing genomic and clinical measures will hinge on the shared vision and mission of team science and education efforts spanning multiple disciplines.

Declaration of interest

WV Bobo has been supported by the National Institutes of Health, the Agency for Healthcare Quality and Research, the National Science Foundation, the Myocarditis Foundation, and the Mayo Foundation for Medical Education and Research; and he has contributed chapters to UpToDate, all of which are unrelated to the present work. AP Athreya has been supported by the National Institutes of Health, National Science Foundation, Blue Gator Foundation, Alzheimer's Association, and the Mayo Foundation for Medical Education and Research. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Funding

This work is based upon work that is supported, in part, by the National Science Foundation (NSF) under award 2041339 and the Mayo Clinic Center for Individualized Medicine.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. GBD. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392 (10159):1789–1858.
2. Bromet E, Andrade LH, Swang I, et al. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. 2011;9(1):90.
3. Murray CJ, Lopez AD. Global morality, disability, and the contribution of risk factors: global burden of disease study. Lancet. 1997;349(9063):1436–1442.
4. World Health Organization (WHO). The global burden of disease: 2004 update. [cited 2021 Mar 9]. Available from: https://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf?ua=1.
5. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA. 2017;317:1517.
6. Liu Q, He H, Yang J, et al. Changes in the global burden of depression from 1990 to 2017: findings from the global burden of disease study. J Psychiatr Res. 2020;126:134–140.
7. Smith K. Mental health: a world of depression. Nature. 2014;515 (7526):181.
8. Wittchen HU, Jacobi F, Rehm J, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. 2011;21(9):655–679.

9. Greenberg PE, Fournier -A-A, Sisitsky T, et al. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry. 2015;76(2):155–162.
10. Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nat Rev Dis Primers. 2016;2(1):160–165.
11. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006;163 (11):1905–1917.
12. Zimmerman M, McGlinchey JB, Posternak MA, et al. How should remission from depression be defined? The depressed patient's perspective. Am J Psychiatry 2006;163:148–150.
13. Blier P. Optimal use of antidepressants: when to act? J Psychiatry Neurosci. 2009;34(1):80.
14. Perlis RH. Abandoning personalization to get to precision in the pharmacotherapy of depression. World Psychiatry. 2016;15 (3):228–235.
15. Gillett G, Tomlinson A, Efthimiou O, et al. Predicting treatment effects in unipolar depression: a meta-review. Pharmacol Ther. 2020;212:107557.
16. Fava M, Uebelacker LA, Alpert JE, et al. Major depressive subtypes and treatment response. Biol Psychiatry. 1997;42(7):568–576.
17. Parker G, Wilhelm K, Mitchell P, et al. Subtyping depression: testing algorithms and identification of a tiered model. J Nerv Ment Dis. 1999;187(10):610–617.
18. Arnow BA, Blasey C, Williams LM, et al. Depression subtypes in predicting antidepressant response: a report from the iSPOT-D trial. Am J Psychiatry. 2015;172(8):743–750.
19. Perlman K, Benrimoh D, Israel S, et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J Affect Disord. 2019;243:503–515.
20. Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and a systematic review. J Affect Disord. 2018;241:519–532.
21. Saveanu RV, Nemeroff CB. Etiology of depression: genetic and environmental factors. Psychiatr Clin North Am. 2012;35 (1):51–71.
22. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Washington, DC, USA: American Psychiatric Association; 2013.
23. Zimmerman M, Ellison W, Young D, et al. How many different ways do patients meeting diagnostic criteria for major depressive disorder? Compr Psychiatry. 2015;56:29–34.
24. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 2013;11(1):126.
25. Milaneschi Y, Lamers F, Peyrot WJ, et al. Polygenetic dissection of major depression clinical heterogeneity. Mol Psychiatry. 2016;21 (4):516–522.
26. Athreya A, Iyer R, Wang L, et al. Integration of machine learning and pharmacogenomic biomarkers for predicting response to antidepressant treatment: can computation all intelligence be used to augment clinical assessments? Pharmacogenomics. 2019;20 (14):983–988.
27. Carter GC, Cantrell RA, Zarotsky V, et al. COMPREHENSIVE REVIEW OF FACTORS IMPLICATED IN THE HETEROGENEITY OF RESPONSE IN DEPRESSION. Depress Anxiety. 2012;29(4):340–354.
28. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Psychology. 2018;14(1):91–118
29. Feczko E, Miranda-Dominguez O, Marr M, et al. The heterogeneity problem: approaches to identify psychiatric subtypes. Trends Cogn Sci. 2019;23(7):584–601.
30. James G, Witten D, Hastie T, et al. An introduction to statistical learning. Vol. 112. 18.
31. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. Value Health. 2019;22(7):808–815.
32. Lin E, Lin C-H, Lane H-Y. Machine learning and deep learning for the pharmacogenomics of antidepressant treatments. Clin Psychopharmacol Neurosci. 2021;19(4):577–588.
33. Tai AMY, Albuquerque A, Carmona NE, et al. Machine learning and big data: implications for disease modeling and a therapeutic discovery in the psychiatry. Artif Intell Med. 2019;99:101704.
34. Bagby RM, Ryder AG, Cristi C. Psychosocial and clinical predictors of response to pharmacotherapy for depression. Journal of psychiatry & neuroscience: JPN. 2002;27(4):250–257.
35. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243–250.
36. Chekroud AM, Gueorguieva R, Krumholz HM, et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry. 2017;74(4):370–378.
37. Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. 2016;78:94–102.
38. Kautzky A, Moller H-J, Dold M, et al. Combining machine learning algorithms for prediction of antidepressant treatment response. Acta Psychiatr Scand. 2021;143(1):36–49.
39. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
40. Tansey KE, Guipponi M, Hu X, et al. Contribution of Kallmann genetic variants to antidepressant response. Biol Psychiatry. 2013;73(7):679–682.
41. Kato M, Serretti A. Review and meta-analysis of antidepressant pharmacogenetic findings in major depressive disorder. Mol Psychiatry. 2010;15(5):473–500.
42. Perlis RH. Pharmacogenomic testing and personalized treatment of depression. Clin Chem. 2014;60(1):53–59.
43. GENDEP Investigators, Mars Investigators, STAR*D Investigators. Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of 3 genome-wide pharmacogenetic studies. Am J Psychiatry. 2013;170(2):207–217.
44. Greden JF, Parikh WV, Rothschild AJ, et al. Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: a large, patient- and rater-blinded, randomized, controlled study. J Psychiatr Res. 2019;111:59–67.
45. Bousman CA, Arandjelovic K, Mancuso SG, et al. Pharmacogenetic tests and depressive symptoms remission: a meta-analysis of randomized controlled trials. Pharmacogenomics. 2019;20(1):37–47.
46. Rosenblat JD, Lee Y, McIntyre RS. The effect of pharmacogenomic testing on response and remission rates in the acute treatment of major depressive disorder: a meta-analysis. J Affect Disord. 2018;241:484–491.
47. Zeier Z, Carpenter LL, Kalin NH, et al. Clinical implementation of pharmacogenetic decision support tools for antidepressant drug prescribing. Am J Psychiatry. 2018;175(9):j873–886.
48. Maciukiewicz M, Marshe VS, Hauschild A-C, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. J Psychiatr Res. 2018;99:62–68.
49. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134(4):382–389.
50. 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. 2021;138:284–290.
    - **This study integrates targeted functional pharmacogenomic biomarkers with clinical/demographic measures and uses multiple machine learning methods with extensive cross-validation to predict therapeutic response to ketamine.**
51. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23(1):56–62.
52. Athreya AP, Neavin D, Carrillo-Roa T, et al., Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. Clin Pharmacol Ther. 2019; 106(4): 855–865.
    - **This study integrates targeted functional pharmacogenomic biomarkers with clinical/demographic measures and machine learning for antidepressant response prediction, with validation in an external dataset and across rating scales.**

53. Gupta M, Neavin D, Liu D, et al. TSPAN5, ERICH3 and selective serotonin reuptake inhibitors in major depressive disorder: pharmacometabolomics-informed pharmacogenomics. Mol Psychiatry. 2016;21(12):1717–1725.
54. Liu D, Ray B, Neavin DR, et al. Beta-defensin 1, aryl hydrocarbon receptor and plasma kynurenine in major depressive disorder: metabolomics-informed genomics. Transl Psychiatry. 2018;8(1):10.
55. Ji Y, Biernacka JM, Hebbring S, et al. Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. Pharmacogenomics J. 2013;13(5):456–463.
56. Rush AJ, Trivdei MH, Ibrahim HM, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry. 2003;54(5):573–583.
57. Biernacka JM, Sangkuhl K, Jenkins G, et al. The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. Transl Psychiatry. 2016;5(4):e553.
58. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006;163(1):28–40.
59. Kautzky A, Baldinger P, Souery D, et al. The combined effect of genetic polymorphisms and clinical parameters on treatment outcome in treatment-resistant depression. Eur Neuropsychopharmacol. 2015;25(4):441–453.
60. Souery D, Oswald P, Massat I, et al. Clinical factors associated with treatment resistance in major depressive disorder: results from a European multi-center study. J Clin Psychiatry. 2007;68(7):1062–1070.
61. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM0IV and ICD-10. J Clin Psychiatry. 1998;59(20):S22–S33.
62. Lin E, Kuo P-H, Liu Y-L, et al. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front Psychiatry. 2018;9:290.
63. Lin E, Kuo P-H, Liu Y-L, et al. Prediction of antidepressant treatment response and remission using an ensemble machine learning framework. Pharmaceuticals. 2020;13(10):305.
64. 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. 2021;11(1):381.
- **This study is one of few that directly compared the performances of numerous supervised learning methods using clinical/demographic measures and pharmacogenomic biomarkers as predictors of antidepressant treatment outcome, with validation in an external dataset.**
65. Fabbri C, Corponi F, Albani D, et al. Pleiotropic genes in psychiatry: calcium channels and stress-related FKBP5 gene in antidepressant resistance. Progr Neuropsychopharmacol Biol Psychiatry. 2018;81:203–210.
66. Iniesta R, Hodgson K, Stahl D, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep. 2018;8(1):5530.
- **Predictive performances of supervised learning models with pharmacogenomic biomarkers and clinical variables as predictors of clinical response to escitalopram or nortriptyline were shown to be highly drug-specific using formal cross-drug sensitivity analyses.**
67. Uher R, Perroud N, Ng MYM, et al. Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. Am J Psychiatry. 2010;167(5):555–564.
68. Shumake J, Mallard JT, McGeary JE, et al. Inclusion of genetic variants in an ensemble of gradient boosting decision trees does not improve the prediction of citalopram treatment response. Sci Rep. 2021;11(1):3780.
69. 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. Progr Neuropsychopharmacol Biol Psychiatry. 2021;104:110036.
70. 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. 2020;52(1):161–171.
71. Fabbri C, Kasper S, Kautzky A, et al. A polygenic predictor of treatment-resistant depression using whole exome sequencing and genome-wide genotyping. Transl Psychiatry. 2020;10(1):50.
- **Whole exome sequencing and genome-wide genotyping were integrated with clinical/demographic predictors in a machine learning algorithm to predict treatment-resistant depression, defined as lack of response to two or more adequate trials of various antidepressants.**
72. Lim S-W, Won -H-H, Kim H, et al. Genetic prediction of antidepressant drug response and nonresponse in Korean patients. PLoS ONE. 2014;9(9):e107098.
73. Yin L, Zhang YY, Zhang X, et al. TPH, SLC6A2, SLC6A3, DRD2 and DRD4 polymorphisms and neuroendocrine factors predict SSRIs treatment outcome in the Chinese population with major depression. Pharmacopsychiatry. 2015;48(3):95–103.
74. Joyce JB, Grant CW, Liu D, et al. Multi-omics drive predictions of response to acute phase combination antidepressant therapy: a machine learning approach with cross-trial replication. Transl Psychiatry. 2021;11(1):513.
- **This study is among the few that used integrative multi-omics (in this case, combining pharmacogenomic and metabolomic biomarkers) with clinical/demographic predictors to predict therapeutic responses to antidepressants using supervised learning models.**
75. Rush AJ, Trivedi MH, Stewart JW, et al. Combining medications to enhance depression outcomes (CO-MED): acute and long term outcomes of a single-blind randomized study. Am J Psychiatry. 2011;168(7):689–701.
76. Bhattacharyya S, Dunlop BW, Mahmoudiandehkordi S, et al. Pilot study of metabolomic clusters as state markers of major depression and outcomes to CBT treatment. Front Neurosci. 2019;13:926.
77. Mahmoudiandehkordi S, Ahmed AT, Bhattacharyya S, et al. Alterations in acylcarnitines, amines, and lipids inform about the mechanism of action of citalopram/escitalopram in major depression. Transl Psychiatry. 2021;11(1):153.
78. Czysz AH, South C, Gadad BS, et al. Can targeted metabolomics predict depression recovery? Results from the CO-MED trial. Transl Psychiatry. 2019;9(1):11.
79. Perna G, Grassi M, Caldirola D, et al. The revolution of personalized psychiatry: will technology make it happen sooner? Psychol Med. 2018;48(5):705–713.
80. Fernandes BS, Williams LM, Steiner J, et al. The new field of 'precision psychiatry. BMC Med. 2017;15(1):80.
81. Ritchie MD, Holzinger ER, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genetics. 2015;16(2):85–97.
82. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3:223–230.
•• **Areview that outlines key differences between machine learning and classical statistics, in terms of mathematical foundations and best uses, as applied to realizing the goal of individualizing the diagnosis and treatment of mental disorders.**
83. Athreya A, Iyer R, Neavin D, et al. Augmentation of physician assessments with multi-omics enhances predictability of drug response: a case study of major depressive disorder. IEEE Comput Intell Mag. 2018;13(3):20–31.
84. The Lancet Respiratory Medicine. Opening the black box of machine learning. Lancet Respir Med. 2018;6(11):801.
85. Chen JH, Asch SM. Machine learning and prediction in medicine – beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507–2509.

86. Porcelli S, Drago A, Fabbri C, et al. Pharmacogenetics of antidepressant response. J Psychiatry Neurosci. 2011;36(2):87–113.
87. Mirza B, Wang W, Wang J, et al. Machine learning and integrative analysis of biomedical big data. Genes (Basel). 2019;10(2):87.
88. Reel PS, Reel S, Pearson E, et al. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021;49:107739.
89. Dantzer, O'Connor JC, Lawson MA, et al. Inflammation-associated depression: from serotonin to kynurenine. Dantzer R, O'Connor JC, Lawson MA, et al. Psychoneuroendocrinology. 2011;36(3):426–436.
90. Lam RW, Milev R, Rotzinger S, et al. Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. BMC Psychiatry. 2016;16(1):105.
91. Menke A, Weber H, Deckert J. Roadmap for routine pharmacogenetic testing in a psychiatric university hospital. Pharmacopsychiatry. 2020;53(4):179–183.
92. Wang L, Scherer SE, Bielinski SJ, et al. Implementation of preemptive DNA sequence-based pharmacogenomic testing across a large academic medical center: the Mayo-Baylor RIGHT 10K Study. Genet Med. 2022;24(5):1062–1072.
93. Fatumo S, Chikowore T, Choudhury A, et al. A roadmap to increase diversity in genomic studies. Nat Med. 2022;28(2):243–250.
94. McIntyre RS, Lee Y, Mansur RB. Treating to target in major depressive disorder: response to remission to functional recovery. CNS Spectr. 2015;20(S1):20–30.
95. Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv. 1995;27(3):326–327.
96. Hofman JM, Watts DJ, Athey S, et al. Integrating explanation and prediction in computational social science. Nature. 2021;595 (7866):181–188.
97. Aradhya S, Nussbaum RL. Genetics in mainstream medicine: finally within grasp to influence healthcare globally. Mol Genet Genomic Med. 2018;6(4):473–480.
98. Cameron LD, Muller C. Psychosocial aspects of genetic testing. Curr Opin Psychiatry. 2009;99(2):218–223.
99. Johansen Taber KA, Dickinson BD. Pharmacogenomic knowledge gaps and educational resource needs among physicians in selected specialties. Pharmacogenomics Pers Med. 2014;7:145–162.
100. Selkirk CG, Weissman SM, Anderson A, et al. Physicians' preparedness for integration of genomic and pharmacogenetic testing into practice within a major healthcare system. Genet Test Mol Biomarkers. 2013;17(3):219–225.
101. Stanek EJ, Sanders CL, Frueh FW. physician awareness and utilization of food and drug administration (FDA)-approved labeling for pharmacogenomic testing information. J Pers Med. 2013;3 (2):111–123.
102. Stanek EJ, Sanders CL, Taber KAJ, et al. Adoption of pharmacogenomic testing by US physicians: results of a nationwide survey. Clin Pharmacol Ther. 2012;91(3):450–458.
103. Salm M, Abbate K, Appelbaum P, et al. Use of genetic tests among neurologists and psychiatrists: knowledge, attitudes, behaviors, and needs for training. J Genet Couns. 2014;23 (2):156–163.
104. Thompson C, Steven PH, Catriona H. Psychiatrist attitudes towards pharmacogenetic testing, direct-to-consumer genetic testing, and integrating genetic counseling into psychiatric patient care. Psychiatry Res. 2015;226(1):68–72.
105. Walden LM, Brandl EJ, Changasi A, et al. Physicians' opinions following pharmacogenetic testing for psychotropic medication. Psychiatry Res. 2015;229(3):913–918.
106. Jacobs M, Pradier MF, McCoy TH, et al. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Transl Psychiatry. 2021;11 (1):108.
107. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–46.
•• **A digestible review highlighting how artificial intelligence/ machine learning technologies can be used to augment human decision-making from the perspectives of clinicians, patients, and healthcare systems.**