Psychological Medicine

cambridge.org/psm

Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis

Review Article

Cite this article: Sajjadian M *et al* (2021). Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychological Medicine* **51**, 2742–2751. https://doi.org/10.1017/S0033291721003871

Received: 17 February 2021
Revised: 13 July 2021
Accepted: 3 September 2021
First published online: 12 October 2021

Keywords:
Machine learning; MDD; meta-analysis; predictive analysis; systematic review; treatment outcome

Author for correspondence:
Rudolf Uher, E-mail: uher@dal.ca

Mehri Sajjadian[1], Raymond W. Lam[2], Roumen Milev[3], Susan Rotzinger[4,5], Benicio N. Frey[6,7], Claudio N. Soares[8], Sagar V. Parikh[9], Jane A. Foster[10], Gustavo Turecki[11], Daniel J. Müller[12,13], Stephen C. Strother[14], Faranak Farzan[15], Sidney H. Kennedy[4,5,16,17] and Rudolf Uher[1]

[1]Department of Psychiatry, Dalhousie University, Halifax, NS, Canada; [2]Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada; [3]Department of Psychiatry and Psychology, Queen's University, Providence Care Hospital, Kingston, ON, Canada; [4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada; [5]Department of Psychiatry, St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada; [6]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada; [7]Mood Disorders Program and Women's Health Concerns Clinic, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada; [8]Department of Psychiatry, Queen's University School of Medicine, Kingston, ON, Canada; [9]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA; [10]Department of Psychiatry & Behavioural Neurosciences, St. Joseph's Healthcare, Hamilton, ON, Canada; [11]Department of Psychiatry, Douglas Institute, McGill University, Montreal, QC, Canada; [12]Campbell Family Mental Health Research Institute, Center for Addiction and Mental Health, Toronto, ON, Canada; [13]Department of Psychiatry, University of Toronto, Toronto, ON, Canada; [14]Baycrest and Department of Medical Biophysics, Rotman Research Center, University of Toronto, Toronto, ON, Canada; [15]eBrain Lab, School of Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC, Canada; [16]Department of Psychiatry, University Health Network, Toronto, ON, Canada and [17]Krembil Research Centre, University Health Network, University of Toronto, Toronto, ON, Canada

Abstract

Background. Multiple treatments are effective for major depressive disorder (MDD), but the outcomes of each treatment vary broadly among individuals. Accurate prediction of outcomes is needed to help select a treatment that is likely to work for a given person. We aim to examine the performance of machine learning methods in delivering replicable predictions of treatment outcomes.

Methods. Of 7732 non-duplicate records identified through literature search, we retained 59 eligible reports and extracted data on sample, treatment, predictors, machine learning method, and treatment outcome prediction. A minimum sample size of 100 and an adequate validation method were used to identify adequate-quality studies. The effects of study features on prediction accuracy were tested with mixed-effects models. Fifty-four of the studies provided accuracy estimates or other estimates that allowed calculation of balanced accuracy of predicting outcomes of treatment.

Results. Eight adequate-quality studies reported a mean accuracy of 0.63 [95% confidence interval (CI) 0.56–0.71], which was significantly lower than a mean accuracy of 0.75 (95% CI 0.72–0.78) in the other 46 studies. Among the adequate-quality studies, accuracies were higher when predicting treatment resistance (0.69) and lower when predicting remission (0.60) or response (0.56). The choice of machine learning method, feature selection, and the ratio of features to individuals were not associated with reported accuracy.

Conclusions. The negative relationship between study quality and prediction accuracy, combined with a lack of independent replication, invites caution when evaluating the potential of machine learning applications for personalizing the treatment of depression.

Introduction

Major depression disorder (MDD) affects 280 million people globally and ranks among the top reasons for disability (World Health Organization, 2021). Dozens of antidepressants, augmentation pharmacological agents, psychological therapies, and brain stimulation procedures are effective for depression (Cipriani et al., 2018; Kennedy et al., 2016; Milev et al., 2016; Parikh et al., 2016), but the efficacy of these treatments varies across individuals. Fewer than half of people with MDD achieve remission with the first treatment (Trivedi et al., 2006). Many have to try multiple treatments before finding an effective one (Malone, 2007; Rush et al., 2006). Each treatment trial takes between 6 and 12 weeks and the delays are associated with the risk of adverse outcomes, including loss of employment and suicide (Al-Harbi, 2012; Crown et al., 2002). If we could predict response to a specific treatment from individual

CAMBRIDGE
UNIVERSITY PRESS

characteristics, we could reduce the duration of depression and improve long-term functional outcomes (Oluboka et al., 2018). Multiple features have been identified as potential predictors of treatment outcomes (Fava, 2009; McGrath et al., 2013; Uher et al., 2012a; Uher, Tansey, Malki, & Perlis, 2012b; Zisook et al., 2007). None of them has been adopted for treatment selection in clinical practice (Perlis, 2016). Reasons for the lack of adoption may be that no single characteristic provides a prediction that is accurate enough to be clinically meaningful or differential prediction of outcomes with alternative treatments (Simon & Perlis, 2010). Since depression is a complex and heterogeneous disorder (Fried, 2017; Wray et al., 2018), multiple features will likely have to be considered in a multivariate model to allow accurate prediction of treatment outcomes (Gillan & Whelan, 2017; Kautzky et al., 2017; Kessler, 2018).

Machine learning is defined as a combination of algorithms that explore how computer systems can learn rules from multiple examples with no need for explicit programming (Samuel, 2000). The computer gradually improves its performance of a task through learning from an increasing amount of data. Machine learning methods can build a model that classifies individuals into predefined categories (e.g. treatment response) or estimates a level of a continuous concept (e.g. degree of reduction in depression severity). The last decade has seen an expansion of machine learning applications in health care, including the prediction of depression treatment outcomes (Lee et al., 2018). In this article, we will synthesize and critically examine the applications of machine learning to depression treatment outcome prediction, evaluate the potential of these methods to inform treatment selection, and propose directions for further research.

Methods

Literature search

We conducted a search of PubMed, Google Scholar, ScienceDirect, and PsychINFO following the PRISMA guidelines (Moher et al., 2009), for articles and reports on MDD, treatment outcomes, and machine learning, published from database inception to 12 October 2020. We used a combination of terms tagging machine learning (statistical learning OR machine learning OR predictive analytics OR deep learning) with terms tagging depression treatment [antidepressant OR depression OR major depressive disorder (MDD)] and its outcomes (treatment outcome OR response OR remission).

Two study authors (M.S. and R.U.) screened the studies and applied the following inclusion criteria: (1) participants with a diagnosis of MDD; (2) clinical assessment with rating scales before and after treatment or historical assessment of treatment resistance; (3) use of a validate machine learning method. The literature search and selection of eligible reports are shown in Fig. 1.

Data extraction

We extracted the size of training, testing, and validation datasets, type of treatment, type and number of predictor variables (clinical variables, demographical variables, treatment history, rating scales for depression, etc.), outcome definition [response, remission, treatment-resistant depression (TRD)], methods used for prediction, missing data, feature selection, validation methods (leave-n-out, k-fold cross-validation, nested cross-validation, holdout or external validation). We recorded the results as accuracy, balanced accuracy, or area under the receiver operating characteristics curve. We transformed available results to the common metrics of balanced accuracy (the average of the reported sensitivity and specificity), which is independent of the proportion of individuals with an outcome of interest.

Study quality assessment

In the absence of a validated quality measure for machine learning studies, we applied minimal requirements for aspects of methodology that have been linked to the replicability of results: sample size and validation procedure. Larger samples are more likely to generate replicable results because they reduce the problems of dimensionality and underfitting (Vabalas, Gowen, Poliakoff, & Casson, 2019). Estimates of the minimal sample size for a machine learning study range from 100 to 300 (Beleites, Neugebauer, Bocklitz, Krafft, & Popp, 2013; Luedtke, Sadikova, & Kessler, 2019). A validation procedure that separates training and testing sets is essential to avoid overfitting. Non-nested cross-validation procedures where feature selection and/or parameter tuning occur in the same loop as predictive accuracy test leads to overfitting (Cawley & Talbot, 2010). Therefore, we required either nested cross-validation or an external validation in a held-out sample with feature selection separated from prediction. We designated studies with a sample size of 100 or more and adequate validation methods as 'adequate-quality'. In addition, a detailed quality assessment following published guidelines (Yusuf et al., 2020) is reported in Supplementary Table S1. All of the adequate-quality papers reported on data sources, data split method, etc., however, none of them reported on the distribution of treatment outcome scores. Moreover, none of the adequate-quality studies used reporting guidelines.

Data analysis

Most studies used more than one machine learning method and reported multiple estimates of predictive accuracy. We used linear mixed-effects models to estimate prediction accuracy and test the effects of methodological features while accounting for the non-independence of multiple estimates with a random effect of the study. In data visualization (e.g. Fig. 2), we plot a single mean estimate of balanced accuracy for each study.

Results

Literature search results

Our literature search retrieved 7732 non-duplicate records. We retained 59 eligible reports that matched our inclusion criteria (Fig. 1). These 59 eligible studies varied in focus, size, and method. The predicted outcomes were remission (19 studies), response (35 studies), and TRD (six studies). The number of individuals ranged from 6 to 36 902 (mean 410, median 115). The predictive variables included demographic, clinical, cognitive, neuroimaging, and molecular genetic variables. The number of features used in prediction ranged from 1 to 4 241 701 (mean 131 660, median 92.5). The ratio of participants to features ranged from 1:1432 to 3690:1.

Accuracy of prediction

Fifty-four of the 59 eligible studies provided accuracy or other estimates (sensitivity and specificity) that allowed the calculation

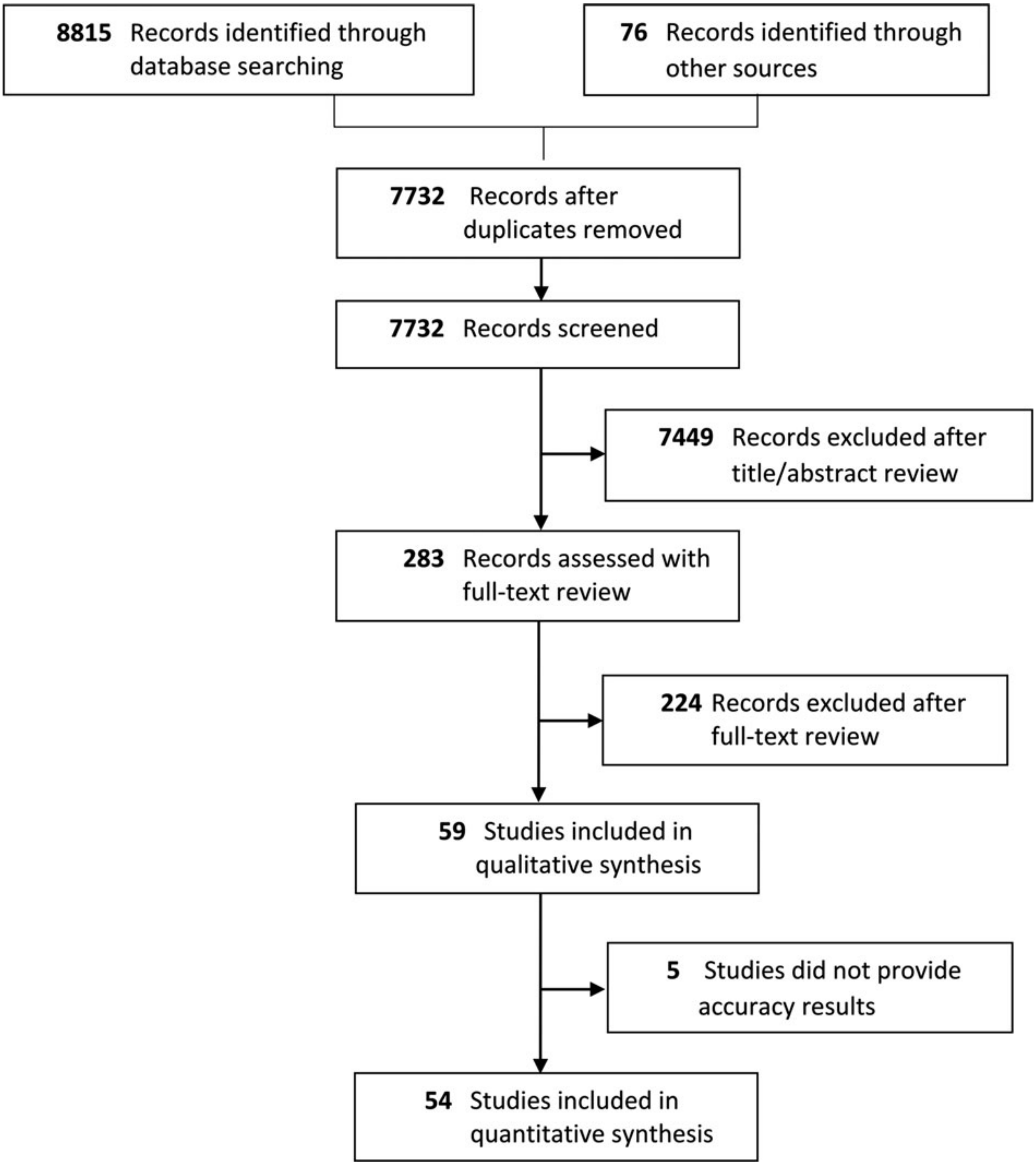
8815 Records identified through database searching
76 Records identified through other sources
7732 Records after duplicates removed
7732 Records screened
7449 Records excluded after title/abstract review
283 Records assessed with full-text review
224 Records excluded after full-text review
59 Studies included in qualitative synthesis
5 Studies did not provide accuracy results
54 Studies included in quantitative synthesis

Fig. 1. Literature search and selection of eligible recoreds for the systematic review and meta-analysis.

of balanced accuracy. Across these studies, we extracted 364 estimates of balanced accuracy, ranging from 0.39 to 1.00. Mean accuracy across estimates within study ranged from 0.48 to 0.91 (mean 0.74, 95% CI 0.71–0.77).

Treatments

The treatments included antidepressant medication (36 studies (61%)), neurostimulation (18 studies (32%)), psychological treatments (four studies (7%)), and other treatments (exercise, psilocybin, blended treatment delivery model; three studies (5%)). Two studies used a combination of two treatment modalities (psychotherapy and antidepressants, neurostimulation, and antidepressant) (Guilloux et al., 2015; Kambeitz et al., 2020).

The outcome of neurostimulation treatment was predicted with greater accuracy (mean 0.79, 95% CI 0.74–0.84) than treatment with antidepressants (mean accuracy 0.70, 95% CI 0.67–0.74) or other treatments (mean accuracy 0.69, 95% CI 0.65–0.73). Most studies predicted outcomes within a single group of participants who received the same treatment. Three studies probed the treatment-specificity of outcome prediction through testing predictive models in groups of participants who received either the same or a different treatment (Chekroud et al., 2016; Iniesta et al., 2018; Kambeitz et al., 2020). In one

https://doi.org/10.1017/S0033291721003871 Published online by Cambridge University Press

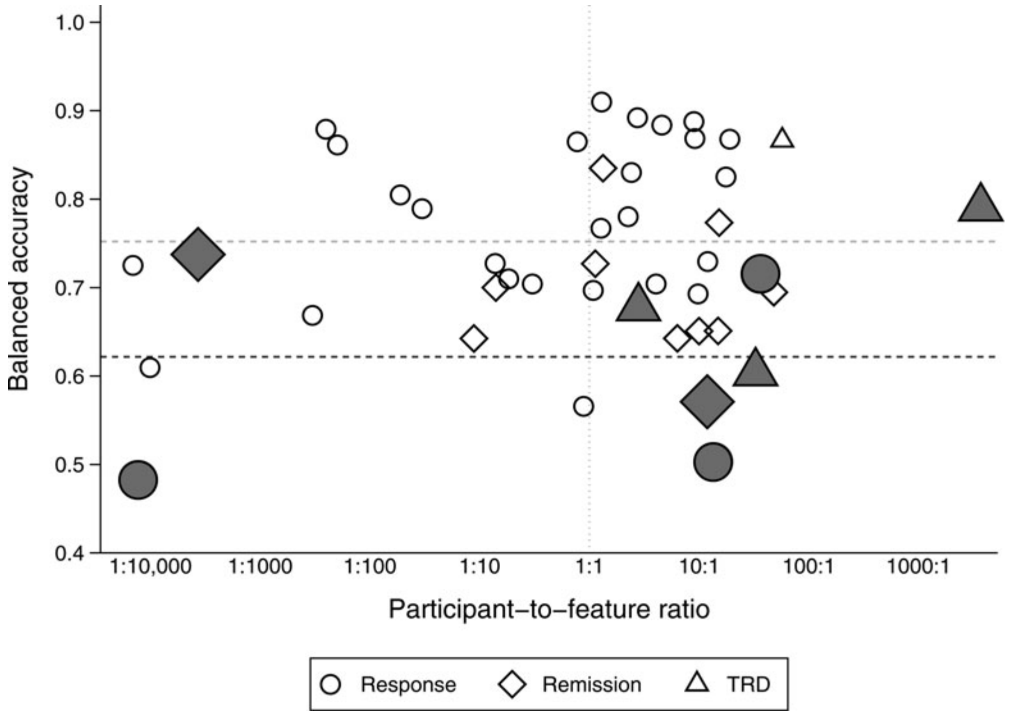

Fig. 2. Balanced accuracy and participant-to-feature ratio in published machine learning studies of outcome prediction in the treatment of MDD. The X-axis plots the ratio of participants to predictive features. Y-axis plots the mean balanced accuracy within each study. Studies predicting response, remission, and TRD are plotted as circles, diamonds, and triangles respectively. Adequate-quality studies are highlighted with large, filled symbols. The dark gray horizontal dashed line shows the mean balanced accuracy of the eight adequate-quality studies. The pale gray horizontal dashed line shows the average balanced accuracy of the other 45 studies.

study, a model based on clinical variables developed in a study of treatment with the antidepressant citalopram significantly predicted outcomes among individuals who received citalopram, but not among those who received a combination of venlafaxine and mirtazapine (Chekroud et al., 2016). Another study used a combination of clinical and genetic variables to derive two models predicting outcomes with escitalopram and nortriptyline respectively, which demonstrated treatment-specificity in a held-out test sample (Iniesta et al., 2018). A third study used clinical and cognitive variables to develop models predicting outcomes of antidepressant and neurostimulation treatment and demonstrated the specificity of predicting escitalopram *vs.* transcranial direct current stimulation (tDCS) outcomes (Kambeitz et al., 2020). In summary, while most studies investigated only one treatment group, three studies suggest that multivariate prediction of outcome is treatment-specific (Chekroud et al., 2016; Iniesta et al., 2018; Kambeitz et al., 2020).

Features contributing to the prediction

The eligible studies used a variety of features as predictors of depression treatment outcomes. Most used neuroimaging (n = 35), followed by clinical and demographic variables ($n = 30$). Relatively few studies used molecular ($n = 8$), and cognitive ($n = 6$) measures. Eighteen studies combined predictors from two modalities: most commonly neuroimaging and clinical ($n = 9$) (Bartlett et al., 2018; Jaworska, De La Salle, Ibrahim, Blier, & Knott, 2019). One study employed a combination of predictors from three modalities of clinical, cognitive, and neuroimaging features (Patel et al., 2015). For further details, please see Supplementary Tables S2 and S4.

There was a significant relationship between feature modality and sample size. Studies that used neuroimaging had small samples (mean 85, median 50 individuals), studies using genetic variables had intermediate sample sizes (mean 307, median 254), and studies using clinical variables had the largest samples (mean 950, median 276). All studies with data on 1000 or more individuals were limited to clinical and demographic variables (Cepeda et al., 2018; Chekroud et al., 2016; Delgadillo & Salas Duhne, 2020; Nie, Vairavan, Narayan, Ye, & Li, 2018; Perlis, 2013).

There was a significant relationship between data modality and reported balanced accuracy. Studies using neuroimaging or genetic data reported significantly higher balanced accuracies ($\beta = 0.13$, 95% CI 0.07–0.18, $p < 0.001$; $\beta = 0.13$, 95% CI 0.08–0.18, $p < 0.001$, respectively) than studies using clinical and demographic variables.

No study tested the added value of neuroimaging to clinical variables within the same sample. One analysis reported improved prediction of treatment outcome with the inclusion of a large number of genetic variables compared to using clinical variables alone (Iniesta, Stahl, & McGuffin, 2016; Iniesta et al., 2018), but a study that used genetic information without clinical features reported prediction not significantly better than chance (Maciukiewicz et al., 2018). Overall, the use of data from multiple modalities was not associated with reported prediction accuracy ($\beta = 0.01$, 95% CI −0.02 to 0.04, $p = 0.641$).

A minority of studies reported on the contribution of specific features. In three analyses of the same large trial sample, initial depression severity and race were among the variables that contributed the most to the predictive models (Chekroud et al., 2016; Nie et al., 2018; Perlis, 2013). Symptoms of reduced interest and activity have also ranked among the most strongly

https://doi.org/10.1017/S0033291721003871 Published online by Cambridge University Press

contributing variables, consistent with the results of univariate analyses (Iniesta et al., 2016, 2018; Uher et al., 2012a, 2012b, 2020).

The number of features used for prediction and the feature-to-observation ratio were unrelated to the reported accuracy of prediction (Fig. 2).

Treatment of missing values

In real-world data, missing values appear due to the loss of participants to follow-up, missed assessments, and intentional or accidental failure to complete items or instruments. Depending on the relationship of missing values to dependent and independent variables of interest, the mechanisms underlying missing values can be classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Many machine learning methods do not support missing values and, consequently, investigators take various options to deal with missing values outside the machine learning algorithms. Ways of handling missing data in machine learning studies include list-wise deletion, replacing with mean/median/mode, predicting the missing values, or using algorithms that support missing values imputation, such as k-nearest neighborhood and random forest. With an increasing number of features, the proportion of individuals with missing data points increases, leading to the loss of a substantial part of the sample at the cost of reduced power. Besides, the deletion of observations with missing values reduces external validity unless the data are MCAR. Imputation of missing values with the prediction by machine learning methods performs better than replacing with mean/median/mode, but care has to be taken to completely separate the imputation between training and testing sets (Bertsimas, Pawlowski, & Zhuo, 2018; Schmitt, 2015; Zhang, 2016). Among the eligible articles, the majority of studies did not address the handling of missing values (43 studies) and other studies used case-wise deletion or mean/mode imputation. Only one study used a machine-learning-based method of handling missing data, the bagged tree imputation (Iniesta et al., 2018). Supplementary Tables S3 and S4 provide detailed information on the treatment of missing values.

Feature selection

Feature selection helps avoid the curse of dimensionality and reduces training time by decreasing the number of features. It also provides information on feature importance and increases generalizability by reducing overfitting (Bermingham et al., 2015; James, Witten, Hastie, & Tibishirani, 2013). Feature selection methods are divided into three main classes: wrapper, filter, and embedded (Guyon, Elisseeff, & Kaelbling, 2003). Wrapper methods employ a predictive model, including the interactions between variables. However, these methods risk overfitting if the number of observations is small and is computationally intensive if the number of variables is large. Filter methods are efficient in calculation time, but often select redundant variables as they do not consider the relationship between features. Embedded methods combine the advantages of wrapper and filter methods. Irrespective of which feature selection method is used, it must be implemented in the training set only to avoid overfitting. Of the 59 included studies, the majority ($n = 33$) did not use any feature selection. We found no relationship between the feature selection method and reported prediction accuracy. For details, please see Supplementary Figs S1 and S2.

Choice of machine learning method

The majority of included studies reported a single machine learning method, but 10 compared multiple methods (Supplementary Fig. S6). The most used methods were regression-based models and support vector machines (Supplementary Fig. S7). Only one study used a deep learning method to predict treatment response (Lin et al., 2018). Within a study, the various methods often gave moderately consistent estimates of balanced accuracy (intraclass correlation 0.62, 95% CI 0.51–0.73). Across the included studies, we found no systematic relationship between the type of machine learning method and the balanced accuracy (Supplementary Tables S4 and S5).

Validation procedures

Validation procedures in machine learning assess the performance of the classification model and its stability across data sets. This is typically achieved by repeatedly dividing the available observations into multiple non-overlapping training and testing sets, an approach known as cross-validation. Details of cross-validation determine the stability of results and the degree of protection against overfitting. Methods with a large overlap between training datasets (e.g. leave-n-out cross-validation) are known to provide less stable estimates than methods with random subsampling (e.g. k-fold cross-validation) whereas the former is less-biased than the latter assuming all other factors are controlled i.e. this is a bias-variance trade-off (Kuhn & Johnson, 2013). Overfitting will occur if the imputation of missing values or feature selection is performed with the entire dataset because the information from the testing set is used in feature selection. Significant overfitting also occurs when feature selection is performed in the same cycle of cross-validation as parameter tuning, because of information leakage between feature selection and parameter tuning (Cawley & Talbot, 2010). Nested cross-validation offers adequate protection against overfitting through separating feature selection from model parameter tuning of the inner and outer cross-validation loops. Another method that offers an adequate test of generalizability is holdout validation, which uses an additional 'unseen' testing dataset that was not used in any way in the model development. Of the 59 eligible studies, 18 reported described validation methods with adequate separation of training and testing sets, including nested cross-validation, external validation in a holdout and/or a separately collected test dataset. The remaining 41 studies used validation methods that may not adequately protect against overfittings (Supplementary Table S4 and Fig. S8).

Study quality and the accuracy of prediction

We defined study quality as a combination of an adequate sample size of 100 or more observations and an adequate validation method with complete separation of training and testing sets at all stages including feature selection (e.g. nested cross-validation or external validation). Of the 59 included studies, 26 had 100 or more participants and 17 reported adequate validation methods. The eight studies that had more than 100 participants and reported adequate validation methods were designated as adequate-quality studies (Table 1, Table 2, and Supplementary

https://doi.org/10.1017/S0033291721003871 Published online by Cambridge University Press

Table 1. Methods for construction of the machine learning model of the eight adequate-quality papers

Study reference	Predictor type	Max *n* of predictors used in the model	Prediction method	Validation method	Variable selection	Additional methods	Missing data imputation	Outcome	Treatment procedure
Athreya et al. (2019)	Pharmacogenomic, clinical	7	Random forest	Nested cross-validation (inner 10-fold cross-validation, outer 5-fold cross-validation), external validation		Clustering	No missing data	Response	Anti-depressant
Cepeda et al. (2018)	Treatment history, administrative	10	Decision tree	External validation, cross-validation				Resistance	Anti-depressant
Chekroud et al. (2016)	Clinical	25	Gradient boosting machine	10-fold cross-validation, external validation	Elastic net model		Including only patients without missing observations	Remission	Anti-depressant
Etkin et al. (2015)	Cognitive emotional biomarkers, clinical, demographical		Logistic regression	Leave-one-out cross-validation on bootstrap subsample, external validation		Linear Discriminant Analysis (LDA)	Were excluded	Remission	Anti-depressant
Iniesta et al. (2018)	Genetic, clinical	20	Elastic net regularized regression	5-fold cross-validation and holdout external replication	CAT score		Bagged tree nonparametric method	Remission	Anti-depressant
Maciukiewicz et al. (2018)	Genotype	38	Support-vector machine (SVM), classification and regression trees (CART)	Nested cross-validation (inner 10-fold cross-validation, outer 5-fold cross-validation)	Logistic regression, lasso regression		Excluded for predictors, for outcome LOCF was used	Response	Anti-depressant
Nie et al. (2018)	Clinical and demographic	700	Random forest, Gradient boosting decision tree, XGBoost, l2 penalized logistic regression, elastic net	10-fold cross-validation, external validation	k-means clustering followed by χ^2 test, elastic net			Resistance	Anti-depressant
Perlis (2013)	Clinical, demographic	15	Logistic regression, random forest, Naive Bayes classifier, SVM	10-fold cross-validation, external validation	Cross-validation in logistic regression model		Mean and mode imputation	Resistance	Anti-depressant

Table 2. Description summary of the eight adequate-quality papers

Study reference	Outcome	Outcome instrument	Treatment duration	Internal				External			
				Accuracy	Balanced accuracy	AUC	Max *n* of subjects	Accuracy	Balanced accuracy	AUC	Max n of subjects
Athreya et al. (2019)	Response	QIDS-C, HDRS	8 weeks	0.80	0.81	0.83	254	0.72	0.71		285
Athreya et al. (2019)	Remission	QIDS-C, HDRS	8 weeks	0.78	0.78	0.86	144	0.75	0.74		182
Cepeda et al. (2018)	Resistance				0.79	0.81	36 902			0.79	9069
Chekroud et al. (2016)	Remission	QIDS-SR16	12 weeks	0.65	0.65		1949	0.57	0.57		151
Etkin et al. (2015)	Remission	HRSD, QIDS-SR16	8 weeks	0.54	0.54		175	0.50			
Etkin et al. (2015)	Response	HRSD, QIDS-SR16	8 weeks	0.54	0.53		175	0.55			
Iniesta et al. (2018)	Remission	HRSD	12 weeks		0.70	0.82	143		0.74	0.77	150
Maciukiewicz et al. (2018)	Response	MADRS	8 weeks	0.61	0.46						
Maciukiewicz et al. (2018)	Remission	MADRS	8 weeks	0.50	0.50		149				37
Nie et al. (2018)	Resistance	QIDS-C16	6–24 weeks				1964	0.71	0.68	0.74	490
Perlis (2013)	Resistance	QIDS-SR	24 weeks	0.69		0.71	1571	0.67	0.61	0.70	523

Table S4). The adequate-quality designation was significantly negatively related to reported accuracy ($b = -0.05$, 95% CI −0.10 to −0.004, $p = 0.035$). Among the adequate-quality studies, the mean balanced accuracy was 0.63 (95% CI 0.56–0.71). Among the remaining 46 studies, the mean balanced accuracy was 0.75 (95% CI 0.72–0.78). The difference in accuracy between adequate-quality and other studies was primarily driven by sample size. The 33 studies with samples smaller than 100 reported a mean balanced accuracy of 0.76 (95% CI 0.73–0.80). The 21 studies with samples of 100 or greater reported a mean balanced accuracy of 0.68 (0.63–0.72). Sample size greater than 100 was significantly negatively related to reported accuracy ($b =-0.05$; 95% CI −0.08 to −0.01, $p = 0.005$). The relationship between adequate validation method and reported balanced accuracy was not significant (−0.02, 95% CI −0.07 to 0.03, $p = 0.469$). Moreover, the adequate-quality studies reported the following range of accuracy for each depression treatment outcome (the confidence intervals of these estimates are relatively broad because of the small number of contributing studies):

(a) response with mean balanced accuracy 0.56 (95% CI 0.43–0.68) based on 17 estimates from three studies;
(b) remission with mean balanced accuracy 0.60 (95% CI 0.51–0.70) based on 16 estimates from five studies;
(c) treatment resistance with mean balanced accuracy 0.69 (95% CI 0.60–0.77) based on 26 estimates from three studies.

Replicability of classification

The likelihood that a prediction will generalize to individuals who were not included in model derivation can be inferred from differences in prediction accuracy between internal cross-validation and external validation or from independent replication in new samples. Only five studies reported accuracy from both internal validation and external validation (Athreya et al., 2019; Browning et al., 2019; Chekroud et al., 2016; Crane et al., 2017; Guilloux et al., 2015). In these studies, the mean balanced accuracy in internal validation was 0.77 and the mean balanced accuracy in external validation was 0.69. The relatively small internal–external drop in accuracy would suggest adequate generalizability, but only a small minority of studies reported relevant data. The preferred way to assess generalizability is independent replication. We found only one attempt at replication in the published literature. One study (Browning et al., 2021) replicated previous work by the same authors predicting antidepressant treatment outcome from measures of symptoms and attentional bias after 1 week of treatment (Browning et al., 2019). The prediction in replication was statistically significant, but the accuracy of the prediction was reduced from 0.80 in the first study to 0.67 in replication (Browning et al., 2019, 2021).

There is an available benchmark dataset that researchers can use to test the generalizability of their algorithms, for example, the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) (Sinyor, Schaffer, & Levitt, 2010).

Discussion

This review synthesizes the rapidly expanding literature on the implementation of machine learning to predict treatment outcomes in depression. Some studies reported promising results, including increased prediction accuracy with the inclusion of multi-modal data, treatment-specific predictions, and positive results from external validation data sets. However, a pattern observed across studies suggests that smaller studies and studies using inadequate validation methods tend to report higher predictive accuracy. This systematic relationship between method and result, coupled with a lack of independent replication, suggests caution in interpreting existing results and the need for careful methodological development.

Several studies suggest that it is possible to derive a multivariate predictive model that is both replicable and treatment-specific. A model developed in a study of nearly 2000 participants based on demographic and clinical variables significantly predicted outcomes in an independent sample of 151 individuals from a study using similar treatment and assessment procedures (Chekroud et al., 2016). While the accuracy of less than 0.60 may not be sufficient for clinical application (Chekroud et al., 2016), other studies suggest that prediction may be improved if data from more modalities are included. In a study of 280 individuals randomly allocated to one of two antidepressant medications, a combination of clinical and molecular genetic features allowed the development of drug-specific prediction models that replicated in a held-out sample of 150 individuals with prediction accuracy over 0.70 (Iniesta et al., 2018). In both studies, the algorithms predicted outcomes in individuals treated with the same type of antidepressant but not among individuals treated with a different type of antidepressant than was used in model development (Chekroud et al., 2016; Iniesta et al., 2018). These promising results of two adequate-quality studies suggest that treatment-specific prediction can be achieved and may be applied to a personalized selection of treatment (Kessler, 2018). However, only a minority of reviewed studies included multiple treatments, limiting the application of results to personalized treatment selection.

While the results of individual studies may be promising, it is important to examine patterns in the literature and consistency across studies. Notably, the machine learning method, feature selection, or features-to-observations ratio were not associated with the reported prediction accuracy. Instead, the sample size and validation design proved essential to the understanding of differences among published studies. We found that some of the highest accuracy estimates had been reported from studies with fewer than 100 participants and/or studies using methods prone to overfitting. While an individual study with fewer than 100 participants may well achieve replicable results, a systematic relationship between study size or quality, and the strength of reported results may indicate bias. The distribution of study size and quality may be partly a result of an early stage in the applications of machine learning methodology to clinical problems and lack of access to large datasets. The largest available datasets are limited to demographic and clinical data (Cepeda et al., 2018; Chekroud et al., 2016; Delgadillo & Salas Duhne, 2020; Nie et al., 2018; Perlis, 2013). Therefore, it is not possible to separate the effect of bias due to low study quality from the potential advantages of additional data modalities. In the next decade, it will be essential to establish large datasets with optimized multimodal assessments that will allow examining the contribution of biomarkers to prediction with adequate methodology.

The success of any machine learning model is defined by its ability to generalize and replicate on a truly independent sample. In recent years, the inability to reproduce the results of many studies has turned into a growing concern among researchers (Baker, 2016). In this context, it is worrying that no full independent replication attempt has been reported for machine learning prediction of depression treatment outcomes. The findings of the present review should make replication a priority for the field of depression treatment outcome prediction.

Criteria for the applicability of machine learning approaches in healthcare

The recent years have seen rapid growth in the publications of studies using machine learning approaches in the prediction of treatment outcomes. However, the methodological rigor of these studies is variable. The present review raises a concern that highly optimistic results might be correlated with insufficient scrutiny of machine learning procedures. The applicability of machine learning algorithms in healthcare will depend on multiple factors, including predictive performance, robustness in calibration across a variety of samples, and proof of an impact on relevant outcomes in practice. Tutorials on how to develop an efficient and reliable machine learning algorithm are now available (Faes et al., 2020; Tohka & van Gils, 2021). In addition, the essential role of using an external validation dataset to prevent overfitting in high-dimensional classification algorithms should be taken into consideration (Park & Han, 2018). Consensus criteria and checklists are now available that allow assessing the adequacy of predictive model development and its applicability (Scott, Carter, & Coiera, 2021; Vollmer et al., 2020).

Future directions

The next decade is expected to see an expansion in open data sharing. Coupled with mature machine learning methodology, the availability of large samples with multimodal measurements will allow separating potential information advantage of adding objective measurement modalities, such as neuroimaging, from publication bias. New data collection in large samples with multiple alternative treatments will improve the clinical applicability of results. Replicability and generalizability are essential features of clinical research and prerequisite to implementation. External validation of a predictive algorithm in a sample that was not available at the time of model development is needed to prove that a machine learning prediction model is reproducible and generalizable.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291721003871.

Acknowledgements. This work was conducted as part of the Canadian Biomarker Integration Network in Depression (CAN-BIND), an Integrated Discovery Program, supported by the Ontario Brain Institute.

Financial support. This work has been funded by the Ontario Brain Institute, which is an independent non-profit corporation, funded partially by the Ontario Government. The opinions, results, and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred. Additional funding was provided by the Canadian Institutes of Health Research (CIHR), Lundbeck, Bristol-Myers Squibb, and Servier. Funding and/or in-kind support was also provided by the investigators' universities and academic institutions. Dr Uher has received additional support from the Canada Research Chairs Program (award number 231397), the Canadian Institutes of Health Research (Grant reference number 148394), and the Dalhousie Medical Research Foundation.

Conflict of interest. Drs. Sajjadian, Foster, Turecki, Farzan, and Uher report no conflict of interest. Dr Lam has received honoraria or research funds from Allergan, Asia-Pacific Economic Cooperation, BC Leading Edge Foundation, CIHR, CANMAT, Canadian Psychiatric Association, Hansoh, Healthy Minds Canada, Janssen, Lundbeck, Lundbeck Institute, MITACS, Ontario

Brain Institute, Otsuka, Pfizer, St. Jude Medical, University Health Network Foundation, and VGH-UBCH Foundation. Dr Milev has received honoraria or research funds from Ontario Brain Institute, Allergan, Janssen, Lallemand, Kye, Lundbeck, Nubiyota, Otsuka, Pfizer, and Sunovion. Dr Rotzinger holds a patent 'Teneurin C-Terminal Associated Peptides (TCAP) and methods and uses thereof. Inventors: David Lovejoy, R.B. Chewpoy, Dalia Barsyte, Susan Rotzinger.' Dr Frey had a research grant from Pfizer. Dr Müller is funded by a CIHR operating grant 428404. Dr Soares has acted as a consultant for Servier, Sunovion, Lundbeck, and Otsuka. Dr Parikh has received honoraria or research funds from Assurex, Takeda, Janssen, Mensante, Aifred, Sage. Dr Strother is the Chief Scientific Officer of ADMdx, Inc., which receives NIH funding. Dr Kennedy has received honoraria or research funds from Abbott, Alkermes, Allergan, BMS, Brain Canada, CIHR, Janssen, Lundbeck, Lundbeck Institute, Ontario Brain Institute, Ontario Research Fund, Otsuka, Pfizer, Servier, Sunovion, Xian-Janssen, and Field Trip Health.

References

Al-Harbi, K. S. (2012). Treatment-resistant depression: Therapeutic trends, challenges, and future directions. *Patient Preference and Adherence*, *6*, 369–388. https://doi.org/10.2147/PPA.S29716.

Athreya, A. P., Neavin, D., Carrillo-Roa, T., Skime, M., Biernacka, J., Frye, M. A., … Bobo, W. V. (2019). Pharmacogenomics-driven prediction of antidepressant treatment outcomes: A machine-learning approach with multi-trial replication. *Clinical Pharmacology and Therapeutics*, *106*(4), 855–865. https://doi.org/10.1002/cpt.1482.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533* (7604), 452–454. https://doi.org/10.1038/533452A.

Bartlett, E. A., DeLorenzo, C., Sharma, P., Yang, J., Zhang, M., Petkova, E., … Parsey, R. V. (2018). Pretreatment and early-treatment cortical thickness is associated with SSRI treatment response in major depressive disorder. *Neuropsychopharmacology*, *43*(11), 2221–2230. https://doi.org/10.1038/s41386-018-0122-9.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., … Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, *5*, 10312. https://doi.org/10.1038/srep10312.

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, *18*(196), 1–39.

Browning, M., Bilderbeck, A. C., Dias, R., Dourish, C. T., Kingslake, J., Deckert, J., … Dawson, G. R. (2021). The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): An open-label, randomised controlled trial. *Neuropsychopharmacology*, *46*(7), 1307–1314. https://doi.org/10.1038/s41386-021-00981-z.

Browning, M., Kingslake, J., Dourish, C. T., Goodwin, G. M., Harmer, C. J., & Dawson, G. R. (2019). Predicting treatment response to antidepressant medication using early changes in emotional processing. *European Neuropsychopharmacology*, *29* (1), 66–75. https://doi.org/10.1016/j.euroneuro.2018.11.1102.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Cepeda, M. S., Reps, J., Fife, D., Blacketer, C., Stang, P., & Ryan, P. (2018). Finding treatment-resistant depression in real-world data: How a data-driven approach compares with expert-based heuristics. *Depression and Anxiety*, *35*(3), 220–228. https://doi.org/10.1002/da.22705.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X.

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., … Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *Lancet*, *391*(10128), 1357–1366. https://doi.org/10.1016/s0140-6736(17)32802-7.

Crane, N. A., Jenkins, L. M., Bhaumik, R., Dion, C., Gowins, J. R., Mickey, B. J., … Langenecker, S. A. (2017). Multidimensional prediction of treatment response to antidepressants with cognitive control and functional MRI. *Brain*, *140*(2), 472–486. https://doi.org/10.1093/brain/aww326.

Crown, W. H., Finkelstein, S., Berndt, E. R., Ling, D., Poret, A. W., Rush, A. J., & Russell, J. M. (2002). The impact of treatment-resistant depression on health care utilization and costs. *Journal of Clinical Psychiatry*, *63*(11), 963–971. https://doi.org/10.4088/JCP.v63n1102.

Delgadillo, J., & Salas Duhne, P. G. (2020). Targeted prescription of cognitive-behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476.

Etkin, A., Patenaude, B., Song, Y. J. C., Usherwood, T., Rekshan, W., Schatzberg, A. F., … Williams, L. M. (2015). A cognitive-emotional biomarker for predicting remission with antidepressant medications: A report from the iSPOT-D trial. *Neuropsychopharmacology*, *40*(6), 1332–1342. https://doi.org/10.1038/npp.2014.333.

Faes, L., Liu, X., Wagner, S. K., Fu, D. J., Balaskas, K., Sim, D., … Denniston, A. K. (2020). A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. *Translational Vision Science and Technology*, *9*(2), 7. https://doi.org/10.1167/tvst.9.2.7.

Fava, M. (2009). Partial responders to antidepressant treatment: Switching strategies. *The Journal of Clinical Psychiatry*, *70*(7), e24. https://doi.org/10.4088/JCP.8017br3c.

Fried, E. (2017). Moving forward: How depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics*, *17*(5), 423–425. https://doi.org/10.1080/14737175.2017.1307737.

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42. https://doi.org/10.1016/j.cobeha.2017.07.003.

Guilloux, J. P., Bassi, S., Ding, Y., Walsh, C., Turecki, G., Tseng, G., … Sibille, E. (2015). Testing the predictive value of peripheral gene expression for nonremission following citalopram treatment for major depression. *Neuropsychopharmacology*, *40*(3), 701–710. https://doi.org/10.1038/npp.2014.226.

Guyon, I., Elisseeff, A., & Kaelbling, L. P. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(7-8), 1157–1182. https://doi.org/10.1162/153244303322753616.

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., … Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, *8*(1), 5530. https://doi.org/10.1038/s41598-018-23584-z.

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455–2465. https://doi.org/10.1017/S0033291716001367.

James, G., Witten, D., Hastie, T., & Tibishirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Jaworska, N., De La Salle, S., Ibrahim, M. H., Blier, P., & Knott, V. (2019). Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. *Frontiers in Psychiatry*, *10*, 768. https://doi.org/10.3389/fpsyt.2018.00768.

Kambeitz, J., Goerigk, S., Gattaz, W., Falkai, P., Benseñor, I. M., Lotufo, P. A., … Brunoni, A. R. (2020). Clinical patterns differentially predict response to transcranial direct current stimulation (tDCS) and escitalopram in major depression: A machine learning analysis of the ELECT-TDCS study. *Journal of Affective Disorders*, *265*, 460–467. https://doi.org/10.1016/j.jad.2020.01.118.

Kautzky, A., Baldinger-Melich, P., Kranz, G. S., Vanicek, T., Souery, D., Montgomery, S., … Kasper, S. (2017). A new prediction model for evaluating treatment-resistant depression. *Journal of Clinical Psychiatry*, *78*(2), 215–222. https://doi.org/10.4088/JCP.15m10381.

Kennedy, S. H., Lam, R. W., McIntyre, R. S., Tourjman, S. V., Bhat, V., Blier, P., … Uher, R. (2016). Canadian network for mood and anxiety treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 3. Pharmacological treatments. *Canadian Journal of Psychiatry*, *61*(9), 540–560. https://doi.org/10.1177/0706743716659417.

https://doi.org/10.1017/S0033291721003871 Published online by Cambridge University Press

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, *31*(1), 32–39. https://doi.org/10.1097/YCO.0000000000000377.

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. In *Applied predictive modeling* (pp. 61–89). New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3.

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., … McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073.

Lin, E., Kuo, P. H., Liu, Y. L., Yu, Y. W., Yang, A. C., & Tsai, S. J. (2018). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in Psychiatry*, *9*, 290. https://doi.org/10.3389/fpsyt.2018.00290.

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445–461. https://doi.org/10.1177/2167702618815466.

Maciukiewicz, M., Marshe, V. S., Hauschild, A. C., Foster, J. A., Rotzinger, S., Kennedy, J. L., … Geraci, J. (2018). GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of Psychiatric Research*, *99*, 62–68. https://doi.org/10.1016/j.jpsychires.2017.12.009.

Malone, D. C. (2007). A budget-impact and cost-effectiveness model for second-line treatment of major depression. *Journal of Managed Care Pharmacy*, *13*(6 SUPPL. A), S8–18. https://doi.org/10.18553/jmcp.2007.13.s6-a.8.

McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., … Mayberg, H. S. (2013). Toward a neuroimaging treatment selection biomarker for major depressive disorder. *JAMA Psychiatry*, *70*(8), 821–829. https://doi.org/10.1001/jamapsychiatry.2013.143.

Milev, R. V., Giacobbe, P., Kennedy, S. H., Blumberger, D. M., Daskalakis, Z. J., Downar, J., … Ravindran, A. V. (2016). Canadian network for mood and anxiety treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 4. Neurostimulation treatments. *Canadian Journal of Psychiatry*, *61*(9), 561–575. https://doi.org/10.1177/0706743716660033.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STARD and an independent clinical study. *PLoS ONE*, *13*(6), e0197268. https://doi.org/10.1371/journal.pone.0197268.

Oluboka, O. J., Katzman, M. A., Habert, J., McIntosh, D., MacQueen, G. M., Milev, R. V., … Blier, P. (2018). Functional recovery in major depressive disorder: Providing early optimal treatment for the individual patient. *International Journal of Neuropsychopharmacology*, *21*(2), 128–144. https://doi.org/10.1093/ijnp/pyx081.

Parikh, S. V., Quilty, L. C., Ravitz, P., Rosenbluth, M., Pavlova, B., Grigoriadis, S., … Uher, R. (2016). Canadian network for mood and anxiety treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 2. Psychological treatments. *Canadian Journal of Psychiatry*, *61*(9), 524–539. https://doi.org/10.1177/0706743716659418.

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*(3), 800–809. https://doi.org/10.1148/radiol.2017171920.

Patel, M. J., Andreescu, C., Price, J. C., Edelman, K. L., Reynolds, C. F., & Aizenstein, H. J. (2015). Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *International Journal of Geriatric Psychiatry*, *30*(10), 1056–1067. https://doi.org/10.1002/gps.4262.

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*(1), 7–14. https://doi.org/10.1016/j.biopsych.2012.12.007.

Perlis, R. H. (2016). Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry*, *15*(3), 228–235. https://doi.org/10.1002/wps.20345.

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., … Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry*, *163*(11), 1905–1917. https://doi.org/10.1176/ajp.2006.163.11.1905.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *44*(1.2), 207–219. https://doi.org/10.1147/rd.441.0206.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, *6*(1), 1–6. https://doi.org/10.472/2155-6180.1000224.

Scott, I., Carter, S., & Coiera, E. (2021). Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health and Care Informatics*, *28*(1), e100251. https://doi.org/10.1136/bmjhci-2020-100251.

Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: Can we match patients with treatments? *American Journal of Psychiatry*, *167*(12), 1445–1455. https://doi.org/10.1176/appi.ajp.2010.09111680.

Sinyor, M., Schaffer, A., & Levitt, A. (2010). The sequenced treatment alternatives to relieve depression (STAR*D) trial: A review. *Canadian Journal of Psychiatry*, *55*(3), 126–135. https://doi.org/10.1177/070674371005500303.

Tohka, J., & van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, *132*, 104324. https://doi.org/10.1016/j.compbiomed.2021.104324.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., … Fava, M. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *American Journal of Psychiatry*, *163*(1), 28–40. https://doi.org/10.1176/appi.ajp.163.1.28.

Uher, R., Frey, B. N., Quilty, L. C., Rotzinger, S., Blier, P., Foster, J. A., … Kennedy, S. H. (2020). Symptom dimension of interest-activity indicates need for aripiprazole augmentation of escitalopram in major depressive disorder: A CAN-BIND-1 report. *Journal of Clinical Psychiatry*, *81*(4), e1–9. https://doi.org/10.4088/JCP.20m13229.

Uher, R., Perlis, R. H., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., … McGuffin, P. (2012a). Depression symptom dimensions as predictors of antidepressant treatment outcome: Replicable evidence for interest-activity symptoms. *Psychological Medicine*, *42*(5), 967–980. https://doi.org/10.1017/S0033291711001905.

Uher, R., Tansey, K. E., Malki, K., & Perlis, R. H. (2012b). Biomarkers predicting treatment outcome in depression: What is clinically significant? *Pharmacogenomics*, *13*(2), 233–240. https://doi.org/10.2217/pgs.11.161.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., … Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *The BMJ*, *368*, l6927. https://doi.org/10.1136/bmj.l6927.

World Health Organization. (2021). Depression. Retrieved from https://www.who.int/news-room/fact-sheets/detail/depression.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., … Sullivan, P. F. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*(5), 668–681. https://doi.org/10.1038/s41588-018-0090-3.

Yusuf, M., Atal, I., Li, J., Smith, P., Ravaud, P., Fergie, M., … Selfe, J. (2020). Reporting quality of studies using machine learning models for medical diagnosis: A systematic review. *BMJ Open*, *10*(3), e034568. https://doi.org/10.1136/bmjopen-2019-034568.

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, *4*(1), 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

Zisook, S., Lesser, I., Stewart, J. W., Wisniewski, S. R., Balasubramani, G. K., Fava, M., … Rush, A. J. (2007). Effect of age at onset on the course of major depressive disorder. *American Journal of Psychiatry*, *164*(10), 1539–1546. https://doi.org/10.1176/appi.ajp.2007.06101757.

https://doi.org/10.1017/S0033291721003871 Published online by Cambridge University Press