Molecular Psychiatry
https://doi.org/10.1038/s41380-020-0825-2

REVIEW ARTICLE

Machine learning for genetic prediction of psychiatric disorders: a systematic review

Matthew Bracher-Smith[1] · Karen Crawford[1,2] · Valentina Escott-Price[1,2]

Received: 15 May 2020 / Revised: 9 June 2020 / Accepted: 16 June 2020

Abstract

Machine learning methods have been employed to make predictions in psychiatry from genotypes, with the potential to bring improved prediction of outcomes in psychiatric genetics; however, their current performance is unclear. We aim to systematically review machine learning methods for predicting psychiatric disorders from genetics alone and evaluate their discrimination, bias and implementation. Medline, PsycInfo, Web of Science and Scopus were searched for terms relating to genetics, psychiatric disorders and machine learning, including neural networks, random forests, support vector machines and boosting, on 10 September 2019. Following PRISMA guidelines, articles were screened for inclusion independently by two authors, extracted, and assessed for risk of bias. Overall, 63 full texts were assessed from a pool of 652 abstracts. Data were extracted for 77 models of schizophrenia, bipolar, autism or anorexia across 13 studies. Performance of machine learning methods was highly varied (0.48–0.95 AUC) and differed between schizophrenia (0.54–0.95 AUC), bipolar (0.48–0.65 AUC), autism (0.52–0.81 AUC) and anorexia (0.62–0.69 AUC). This is likely due to the high risk of bias identified in the study designs and analysis for reported results. Choices for predictor selection, hyperparameter search and validation methodology, and viewing of the test set during training were common causes of high risk of bias in analysis. Key steps in model development and validation were frequently not performed or unreported. Comparison of discrimination across studies was constrained by heterogeneity of predictors, outcome and measurement, in addition to sample overlap within and across studies. Given widespread high risk of bias and the small number of studies identified, it is important to ensure established analysis methods are adopted. We emphasise best practices in methodology and reporting for improving future studies.

Introduction

Machine learning (ML) represents a contrasting approach to traditional methods for genetic prediction. It has increased in popularity in recent years following breakthroughs in deep learning [1–4], and the scaling-up of datasets and computing power. The ability to function in high dimensions and detect interactions between loci [5] without assuming additivity makes such methods an attractive option in statistical genetics, where the effects of myriad factors on an outcome is difficult to pre-specify. Calls to address the complexity of disorders like schizophrenia with ML have also become more frequent [6–8]. However, the predictive performance of ML methods in psychiatric genetics is unclear, and a recent review of clinical prediction models across various outcomes and predictors found them to be no more accurate than logistic regression (LR) [9]; it is therefore timely to review their predictive performance in psychiatry.

Genome-wide association studies (GWAS), genetic prediction and psychiatry have each been reviewed with respect to ML [10–16]. Recently, single-nucleotide polymorphism (SNP)-based prediction has been reviewed across diseases [17]. However, psychiatry presents a distinct problem from somatic and neurological diseases as a result of

Supplementary information The online version of this article (https://doi.org/10.1038/s41380-020-0825-2) contains supplementary material, which is available to authorised users.

✉ Valentina Escott-Price
escottpricev@cardiff.ac.uk

[1] MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

[2] Dementia Research Institute, School of Medicine, Cardiff University, Cardiff, UK

Published online: 26 June 2020

genetic correlation between disorders [18] and the risk of class mislabelling due to biological heterogeneity that may underlie symptom-based diagnoses [19].

We systematically reviewed literature related to the question: what is the ability of ML methods to predict psychiatric disorders using only genetic data? We report discrimination, methodology and potential bias for diagnostic or prognostic models and compare with LR and polygenic risk scores (PRS) where available.

Materials and methods

Search strategy

Medline via Ovid, PsycInfo, Web of Science and Scopus were searched for journal articles matching terms for ML, psychiatric disorders and genetics on 10 September 2019. Searches were broad, with terms for psychiatric disorders including schizophrenia, bipolar, depression, anxiety, anorexia and bulimia, attention-deficit hyperactivity disorder, obsessive compulsive disorder, Tourette's syndrome or autism. Terms for ML were also wide-ranging, including naive Bayes, *k*-nearest neighbours (*k*-NN), penalised regression, decision trees, random forests, boosting, Bayesian networks, Gaussian processes, support vector machines (SVMs) and neural networks, but excluding regression methods without penalty terms, such as LR. Searches were developed and conducted by MBS and were restricted to English language journal articles on humans, with no limits on search dates. Two authors (MBS, KC) independently reviewed all abstracts for inclusion. Full texts were assessed if either author had chosen to access them and independently screened against inclusion criteria. Where conflicts occurred a third author (VEP) was consulted as an arbiter. An example search for Medline (Ovid) is given in the Supplementary (Table S1).

Inclusion and exclusion criteria

Studies were restricted to cohort, cross-sectional or case–control designs of individuals for binary classification of a single DSM or ICD-recognised psychiatric disorder compared with unaffected individuals, where only genotyping array, exome or whole-genome sequencing data were used as predictors. Studies based solely on gene expression were excluded, but designs which made use of gene expression or functional annotations to inform models of genetic data were accepted. No further restriction was made on participants. Studies were excluded if they only predicted medication response, subgroups within a psychiatric disorder or a psychiatric phenotype secondary to another disease. Studies were also considered ineligible if they had a clear primary aim of drawing inference at the expense of prediction, if they developed a novel statistical method or only made use of unsupervised or semi-supervised methods. The review was registered to PROSPERO in advance (registration number CRD42019128820).

Extraction and analysis

A data extraction form was developed through discussion between all authors. Items from the critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist [20] were included as-is or modified. Additional items were included based on expert knowledge and relevance to the review topic, with reference to the genetic risk prediction studies (GRIPS) statement [21] for items pertaining to genetic models (Table S2). The form was piloted with five publications, containing 40 extracted ML models between them, and updated before being applied to all texts.

The discrimination of ML methods was extracted independently by two authors (MBS and KC) as area under the receiver operating characteristic curve (AUC), or *c*-statistic. Model performance measures for classification by accuracy, sensitivity and specificity were also extracted. 95% confidence intervals for validation were estimated for AUC using Newcombe's method [22]. Results were not meta-analysed due to sample overlap, present in at least half of studies (see Table S3), which cannot easily be accounted for in the meta-analysis. Information on participants, predictors and model development and validation were also obtained. LR or PRS models were also extracted when present. Though LR can be considered a ML approach, for the purpose of this review we regard it as a contrasting method due to its widespread use in classic statistical analysis. The presence of LR and PRS as comparators was not made a requirement due to their sparsity in the literature.

Risk of bias (ROB) and applicability were assessed using the prediction model risk of bias assessment tool (PROBAST) [23]. PROBAST consists of 20 questions designed to signal where ROB may be present in either the development or validation of a model across four categories—participants, predictors, outcome and analysis. These include, for instance, questions on how missingness or complexities in study design were handled. Information on handling of population structure, a common confound in genetic association studies, was also extracted to aid ROB assessment. Reporting of the systematic review follows the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines [24]. Extraction and ROB are detailed further in the Supplementary.

Results

Selection

In total, 1241 publications were identified through searches in Ovid Medline, PsycInfo, Scopus and Web of Science which included restrictions to English language journal articles (Fig. S1). After merging and removing duplicates, 652 studies were assessed for inclusion. Of these, 63 full texts were assessed to determine eligibility. Fourteen publications were selected, with two merged as publications included the same models on the same dataset. A final total of 13 studies were selected for inclusion, containing 77 distinct ML models.

Studies

A wide range of ML methods were applied to schizophrenia (7 studies, 47% of models), bipolar disorder (5 studies, 39% of models), autism (3 studies, 10% of models) and anorexia (1 study, 4% of models) (Table 1), with no studies identified for the six remaining disorders. SNPs were the most common source of genetic data. Copy number variants (CNVs) and PRSs were each incorporated in models from a single study, and exome-sequencing data formed the basis of two studies. Datasets typically consisted of publicly available GWAS; potential sample overlap was established for at least seven studies (Table S3). Briefly, three studies [25–27] included controls for the 1958 Birth Cohort [28] or the UK Blood Service [29], four studies included controls from Knowledge Networks [25, 30–32], two studies used a Swedish population-based sample [32, 33], and three studies used the same dataset, or provided a common reference for part of the dataset [25, 30, 31]. The remaining six studies [34–40] either gave unclear information, reported no previous reference for the dataset, or used datasets which appear to be separate from other studies. Where samples overlap, all models included in the review are distinct, using different predictors or modelling approaches. Additional overlap or cryptic relatedness may be present between studies.

Missingness was reported clearly in about half of all studies and models. When reported, it was most commonly handled by imputation after excluding genotypes with high missingness. Studies also reported complete-case analysis and inclusion of missing values in coding of predictors (Table S4).

Machine learning methods

SVMs and neural networks were the most popular, followed by random forests and boosting. SVMs were split roughly equally between using a linear kernel (3 studies, 7 models), a radial basis function (RBF) kernel (3 studies, 6 models) or an unreported kernel (3 studies, 6 models). Authors applying neural networks most commonly used multi-layer perceptrons (3 studies, 6 models), an RBF network (2 studies, 5 models) or restricted Boltzmann machines (RBMs; 1 study, 9 models), with linear networks, convolutional neural networks (CNNs) and embedding layers each used once. Weak learners in boosted models were mainly decision trees, with the exception of a method which combined feature selection with the boosting of RBF-SVMs in AdaBoost [35]. Penalised regression was employed alongside linear and non-linear methods as least absolute shrinkage and selection operator (LASSO; 3 studies, 4 models) or ridge regression (1 study, 2 models). Overall, 51% of all models were implemented in R or WEKA; Matlab and Python were preferred for neural networks (Table S5).

Risk of bias

ROB was assessed for each model within each study (Fig. S2). All models displayed ROB, mostly in relation to participants (study design and inclusion/exclusion criteria), outcome (standardised definition and assessment of outcomes) and analysis. Within-study ROB for participants was due to the use of case–control studies. Predictors were mostly rated to have unclear or low ROB; instances of high ROB were limited to predictors which are unavailable at the point of model use. Outcome definitions or measurements often differed between cases and controls.

Models displayed high ROB during analysis. This was often traced to inappropriate or unjustified handling of missingness and removal of enroled participants prior to analysis, predictor selection using univariable methods and failure to account for overfitting. No studies reported calibration measures. In addition to PROBAST, information on population structure within studies was extracted (Table S6). Most studies did not illustrate genetic ancestry across all observations in the current publication using dimensionality reduction, and none reported any evaluation of the final trained model for bias due to population structure. However, two studies (18% of models) visualised principal components for a subsample or showed a table of reported ancestry for participants [31, 39]. Where ancestry was not addressed in a study, it was most often visualised in a referenced publication (55% of all models). Two studies (13% of models) had no details or references which addressed genetic ancestry.

Across-study ROB was not formally assessed. For schizophrenia, bipolar and autism, studies with smaller numbers of cases in the development set report AUC less often, instead preferring classification metrics such as accuracy, sensitivity and specificity.

Table 1 Overview of studies.

First author (Ref.)	Disorder	Machine learning methods	Data	Models	Comparators
Aguiar-Pulido et al. [34, 36][a]	Schizophrenia	AdaBoost, BFTree, DNTB, decision tables, SVM (kernel not reported), naive Bayes, Bayesian networks, MDR, neural network (RBF, linear, perceptron), evolutionary computation	SNPs	12	
Yang et al. [35]	Schizophrenia	AdaBoost (of SVM (RBF)), SVM (RBF)	SNPs	2	
Pirooznia et al. [25]	Bipolar disorder	Bayesian networks, random forest, neural network (RBF), SVM (kernel not reported)	SNPs	16	PRS, LR
Li et al. [30]	Bipolar disorder, Schizophrenia	LASSO, Ridge, SVM (linear)	SNPs	6	
Engchuan et al. [37]	Autism	Neural network (perceptron), SVM (Linear), random forest, CIF	CNVs	4	
Acikel et al. [31]	Bipolar disorder	MDR, random forest, *k*-NN, naive Bayes	SNPs	5	
Guo et al. [26]	Anorexia nervosa	LASSO, SVM (RBF), GBM	SNPs	3	
Laksshman et al. [38]	Bipolar disorder	Decision tree, random forest, neural network (CNN)	Exomes	3	
Chen et al. [32]	Schizophrenia	Neural network (perceptron)	PRS	4	PRS, LR
Wang et al. [39]	Schizophrenia, Bipolar disorder, Autism	Neural networks (cRBM)	SNPs, gene expression	9	LR
Ghafouri-Fard et al. [40]	Autism	Neural network (with embedding layer)	SNPs	1	
Trakadis et al. [33]	Schizophrenia	LASSO, random forest, SVM (kernel not reported), GBM (XGBoost)	Exomes	4	
Vivian-Griffiths et al. [27]	Schizophrenia	SVM (linear, RBF)	SNPs	8	PRS

BFTree best-first decision tree, *CIF* conditional inference forest, *cRBM* conditional restricted Boltzmann machine, *CNN* convolutional neural network, *DNTB* Decision table naive Bayes, *GBM* gradient boosting machine, *k-NN* *k*-nearest neighbours, *LASSO* least absolute shrinkage and selection operator, *LR* logistic regression, *MDR* multifactor dimensionality reduction, *PRS* polygenic risk score, *RBF* radial basis function, *SVM* support vector machine.

[a] Merged in extraction [34, 36].

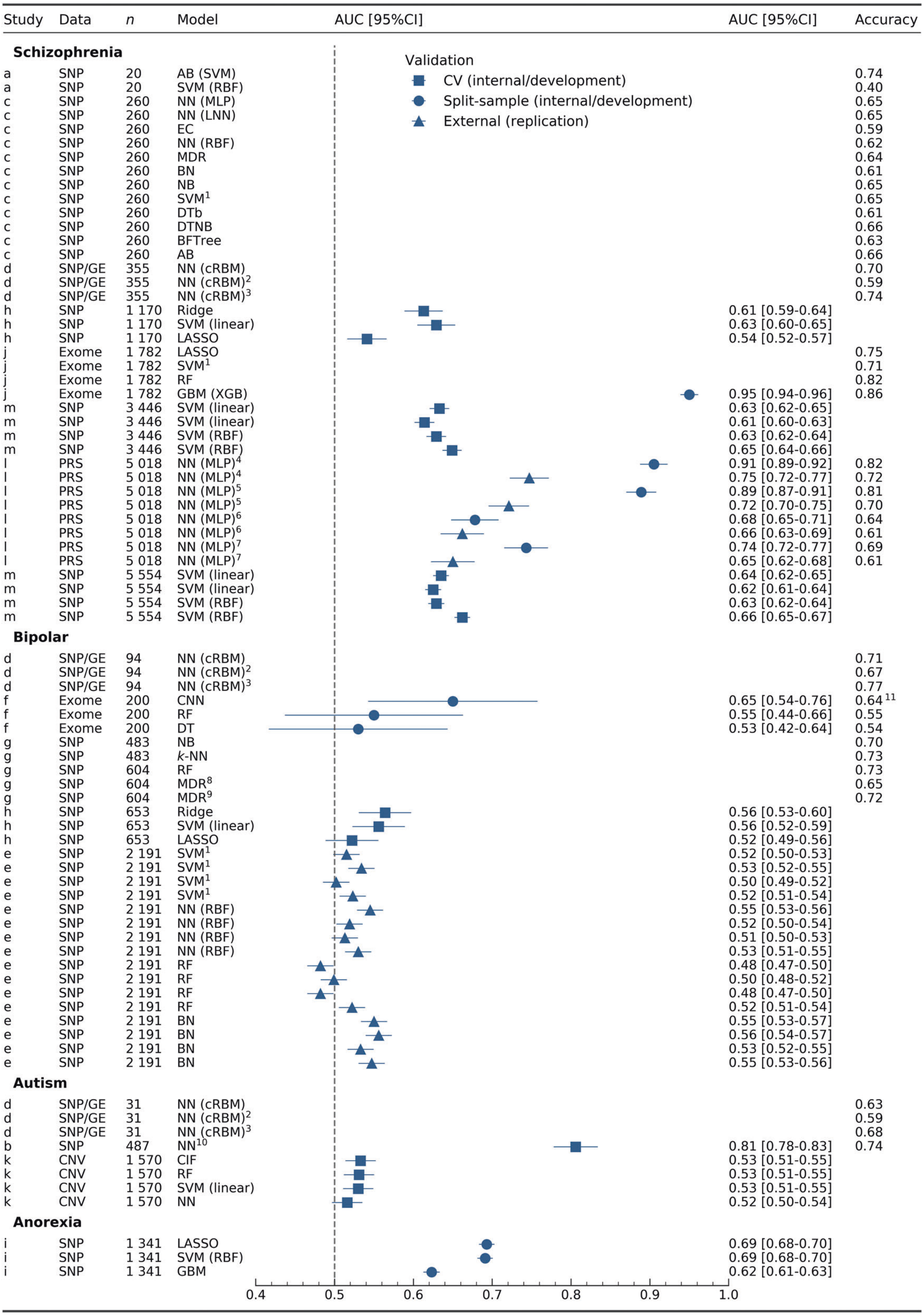
PROBAST encourages assessment of studies for applicability to the review question as this is often narrower than inclusion criteria [23]. Concern was identified for models in three studies [25, 30, 39]. All others demonstrated either low concern or unclear applicability. Reasons for concern were attributable to outcomes which combined closely related disorders, or the use of post-mortem gene expression data, whereas the review question focussed on models of single disorders with potential use in diagnosis or prognosis.

Model performance

Over half of all models assessed discrimination using AUC (58% models). A wide range of classification metrics and measures of model fit were also reported (Table S7), with less than a quarter of models clearly reporting choosing a decision threshold a priori (Table S8).

Around 79% of models, from 12 studies, reported some form of internal validation (Table S9). The majority of these were *k*-fold cross-validation (CV; 57% of all models; 8 studies), a resampling approach which involves testing a model on each of *k* independent partitions of a dataset, every time training on the remaining *k*-1 folds. Tenfold CV was most commonly used, with just below half of all CV models invoking repeats with different random splits. The remainder of studies using internal validation created a random split between training and testing sets (21% of all models; 5 studies), or applied apparent validation, where training and testing are both done on the whole sample [31]. A minority reported external validation (26% of models; 2 studies). Use of internal validation was not reported for 16 models from a single study [25], but for which geographic and temporal external validation was given. External validation was reported for one other study, but with partly overlapping participants between development and validation sets [32].

Model performance varied by choice of statistical method, sample size and number of predictors within studies (Table S10). Discrimination for models of schizophrenia (Fig. 1) was extremely varied (0.541–0.95 AUC), with the highest AUC from exome data using XGBoost (0.95 AUC) [33]. In this study, Trakadis et al. [33] used counts of variants in each gene, after annotation and predictor selection, on participants with part-Finnish or Swedish ancestry [41]. Similarly, high AUC (0.905 AUC) made use of multiple schizophrenia-associated PRS [32]. However, the authors identify the presence of both the development and validation samples in the psychiatric genomics consortium GWAS used to generate the schizophrenia PRS [42], in addition to having overlapping controls between internal validation (model development) and external validation (replication) samples. All other schizophrenia models involved learning from SNPs [27, 30, 34–36], with the exception of Wang et al. [39] where gene expression data from post-mortem samples informed the weights in a conditional RBM trained on genotypes.



Study	Data	n	Model	AUC [95%CI]	Accuracy
Schizophrenia					
a	SNP	20	AB (SVM)		0.74
a	SNP	20	SVM (RBF)		0.40
c	SNP	260	NN (MLP)		0.65
c	SNP	260	NN (LNN)		0.65
c	SNP	260	EC		0.59
c	SNP	260	NN (RBF)		0.62
c	SNP	260	MDR		0.64
c	SNP	260	BN		0.61
c	SNP	260	NB		0.65
c	SNP	260	SVM[1]		0.65
c	SNP	260	DTb		0.61
c	SNP	260	DTNB		0.66
c	SNP	260	BFTree		0.63
c	SNP	260	AB		0.66
d	SNP/GE	355	NN (cRBM)		0.70
d	SNP/GE	355	NN (cRBM)[2]		0.59
d	SNP/GE	355	NN (cRBM)[3]		0.74
h	SNP	1 170	Ridge	0.61 [0.59-0.64]	
h	SNP	1 170	SVM (linear)	0.63 [0.60-0.65]	
h	SNP	1 170	LASSO	0.54 [0.52-0.57]	
j	Exome	1 782	LASSO		0.75
j	Exome	1 782	SVM[1]		0.71
j	Exome	1 782	RF		0.82
j	Exome	1 782	GBM (XGB)	0.95 [0.94-0.96]	0.86
m	SNP	3 446	SVM (linear)	0.63 [0.62-0.65]	
m	SNP	3 446	SVM (linear)	0.61 [0.60-0.63]	
m	SNP	3 446	SVM (RBF)	0.63 [0.62-0.64]	
m	SNP	3 446	SVM (RBF)	0.65 [0.64-0.66]	
l	PRS	5 018	NN (MLP)[4]	0.91 [0.89-0.92]	0.82
l	PRS	5 018	NN (MLP)[4]	0.75 [0.72-0.77]	0.72
l	PRS	5 018	NN (MLP)[5]	0.89 [0.87-0.91]	0.81
l	PRS	5 018	NN (MLP)[5]	0.72 [0.70-0.75]	0.70
l	PRS	5 018	NN (MLP)[6]	0.68 [0.65-0.71]	0.64
l	PRS	5 018	NN (MLP)[6]	0.66 [0.63-0.69]	0.61
l	PRS	5 018	NN (MLP)[7]	0.74 [0.72-0.77]	0.69
l	PRS	5 018	NN (MLP)[7]	0.65 [0.62-0.68]	0.61
m	SNP	5 554	SVM (linear)	0.64 [0.62-0.65]	
m	SNP	5 554	SVM (linear)	0.62 [0.61-0.64]	
m	SNP	5 554	SVM (RBF)	0.63 [0.62-0.64]	
m	SNP	5 554	SVM (RBF)	0.66 [0.65-0.67]	
Bipolar					
d	SNP/GE	94	NN (cRBM)		0.71
d	SNP/GE	94	NN (cRBM)[2]		0.67
d	SNP/GE	94	NN (cRBM)[3]		0.77
f	Exome	200	CNN	0.65 [0.54-0.76]	0.64[11]
f	Exome	200	RF	0.55 [0.44-0.66]	0.55
f	Exome	200	DT	0.53 [0.42-0.64]	0.54
g	SNP	483	NB		0.70
g	SNP	483	*k*-NN		0.73
g	SNP	604	RF		0.73
g	SNP	604	MDR[8]		0.65
g	SNP	604	MDR[9]		0.72
h	SNP	653	Ridge	0.56 [0.53-0.60]	
h	SNP	653	SVM (linear)	0.56 [0.52-0.59]	
h	SNP	653	LASSO	0.52 [0.49-0.56]	
e	SNP	2 191	SVM[1]	0.52 [0.50-0.53]	
e	SNP	2 191	SVM[1]	0.53 [0.52-0.55]	
e	SNP	2 191	SVM[1]	0.50 [0.49-0.52]	
e	SNP	2 191	SVM[1]	0.52 [0.51-0.54]	
e	SNP	2 191	NN (RBF)	0.55 [0.53-0.56]	
e	SNP	2 191	NN (RBF)	0.52 [0.50-0.54]	
e	SNP	2 191	NN (RBF)	0.51 [0.50-0.53]	
e	SNP	2 191	NN (RBF)	0.53 [0.51-0.55]	
e	SNP	2 191	RF	0.48 [0.47-0.50]	
e	SNP	2 191	RF	0.50 [0.48-0.52]	
e	SNP	2 191	RF	0.48 [0.47-0.50]	
e	SNP	2 191	RF	0.52 [0.51-0.54]	
e	SNP	2 191	BN	0.55 [0.53-0.57]	
e	SNP	2 191	BN	0.56 [0.54-0.57]	
e	SNP	2 191	BN	0.53 [0.52-0.55]	
e	SNP	2 191	BN	0.55 [0.53-0.56]	
Autism					
d	SNP/GE	31	NN (cRBM)		0.63
d	SNP/GE	31	NN (cRBM)[2]		0.59
d	SNP/GE	31	NN (cRBM)[3]		0.68
b	SNP	487	NN[10]	0.81 [0.78-0.83]	0.74
k	CNV	1 570	CIF	0.53 [0.51-0.55]	
k	CNV	1 570	RF	0.53 [0.51-0.55]	
k	CNV	1 570	SVM (linear)	0.53 [0.51-0.55]	
k	CNV	1 570	NN	0.52 [0.50-0.54]	
Anorexia					
i	SNP	1 341	LASSO	0.69 [0.68-0.70]	
i	SNP	1 341	SVM (RBF)	0.69 [0.68-0.70]	
i	SNP	1 341	GBM	0.62 [0.61-0.63]	

Predictive ability for bipolar disorder (Fig. 1) was consistently lower than for schizophrenia, frequently overlapping with chance (0.482–0.65 AUC). Models were trained on genotypes, excepting a study [38] using exome data to train a CNN as part of the Critical Assessment of Genome Interpretation (CAGI)

◀ **Fig. 1 Discrimination for all models.** n number of cases in training set. Studies: a [35], b [40], c [34, 36], d [39], e [25], f [38], g [31], h [30], i [26], j [33], k [37], l [32], m [27]. [1]SVM kernel not reported. [2]Modified architecture with intermediate phenotypes in training set only. [3]Modified architecture with intermediate phenotypes for training and test sets. [4,5,6,7]Internal and external validation are shown for study l, where validations for the same model are denoted with the same number. [8]Two-way MDR. [9]Three-way MDR. [10]Neural network embedding layer. [11]Accuracy calculated from confusion matrix. AB AdaBoost, BN Bayesian networks, BFTree best-first tree, CIF conditional inference forest, cRBM conditional restricted Boltzmann machine, CI confidence interval, CNN convolutional neural network, CNV copy number variation, DTb decision tables, DTNB decision table naive Bayes, DT decision tree, EC evolutionary computation, GE gene expression, GBM gradient boosting machine, *k*-NN *k*-nearest neighbours, LASSO least absolute shrinkage and selection operator, LNN linear neural network, MDR multifactor dimensionality reduction, MLP multi-layer perceptron, NB naive Bayes, NN neural network, PRS polygenic risk scores, RBF radial basis function, RF random forests, SNP single-nucleotide polymorphisms, SVM support vector machine, XGB extreme gradient boosting.

competition [43], for which moderate discrimination was achieved (0.65 AUC).

Significantly fewer models were reported for autism (8 models, 3 studies) and anorexia (3 models, 1 study) (Fig. 1). Varying predictive performance was illustrated in autism (0.516–0.806 AUC). High AUC (0.806 AUC) was shown for a single prediction model [40], while models developed with a greater sample size by Engchuan et al. using CNVs were closer to or overlapping with chance (0.516–0.533 AUC) [37]. The only models predicting anorexia nervosa had moderate discriminative ability between cases and controls (0.623–0.693 AUC) [26].

Logistic regression and polygenic risk scores

Three studies reported AUC for either LR (5 models) or PRS (12 models) alongside ML methods. PRS were weighted by summary statistics from a GWAS on the same disorder as the outcome and used as the sole predictor in a LR model. Though discrimination shows some difference between model types, the number of studies for comparison is low and results are clustered by study and type of validation (Fig. S3).

Predictors

Coding of predictors was mostly unclear or unreported (7 studies, 55% of models). Coding was unclear if it was implied through the description of the type of classifier or software but not clearly articulated for the reported study. PRS were continuous [32] while counts of variants-per-gene or genes-per-gene-set were used for exomes and CNVs respectively [33, 37]. SNPs were coded under an additive model, a z-transformation of additive coding, or one-hot encoded (one predictor per genotype at a locus) (Table S11). GWAS summary statistics from external datasets were also used in the selection, weighting or combining of predictors (9 studies, 64% models; Table S12).

Predictor selection was adopted by most (12, 73% of models) and limited to filter-based selection, used prior to modelling, and embedded selection, an integral part of the prediction model (Table S13). The latter involved LASSO regression, or ensembles and hybrids of decision trees and decision tables, in addition to a modified AdaBoost [35]. Filters were based on internal or external univariable association tests (GWAS). Embedded and wrapper-based methods, which typically 'wrap' a model in forward or backward-selection, were both also used prior to any predictive modelling. Modification of predictors using information from the test set was the most common cause of information 'leaking' from the test set to the training set, a source of inflation in performance measures (Table S14).

Sample size

Total sample size was generally low where a single sample had been used, but higher if genotypes from publicly available amalgamated datasets used in a GWAS had been downloaded (median 3486, range 40–11853) (Table S10). Number of events in development followed a similar pattern (median 1341, range 20–5554) as class imbalance was minimal (median 1, range 0.65–2.93, calculated as non-events over events). Around half of studies gave sufficient information to calculate events per variable (median 0.69, range 0.00063–74.6). It could not be calculated where the number of candidate predictors where not reported for models in two studies [25, 39]; approximations are given in the Supplementary where reporting was unclear in a further five studies [26, 32–34, 36, 38] (Table S10).

Hyperparameter search

Hyperparameter search was mostly unreported or unclear (41 models, 9 studies), with some models reported as having been used with default settings. Ambiguous reporting resulted from description of search and tuning for a specific model, with no clarity as to whether these conditions applied to other models in the study. Only 19% of models clearly reported attempting different hyperparameters for the extracted models (Table S15). Studies also report non-standard final hyperparameters, such as uneven batch size in neural networks, or showed good accuracy for a model which is highly sensitive to tuning of crucial hyperparameters, yet few reported tuning (Table S16). It is therefore likely that most studies evaluated several hyperparameter choices but did not report this.

Discussion

All studies displayed high ROB in model development and validation with infrequent reporting of standard modelling steps. Performance measures consequently demonstrated a wide range of abilities to discriminate between cases and controls (0.482–0.95 AUC). These are likely optimistic owing to the high ROB identified through PROBAST and unaddressed sample overlap and population structure, as two studies showing the highest AUCs left these issues unresolved [32, 33]. Though potential bias and effective sample size limit overall interpretation of discrimination, low standards of model development, validation and reporting are a clear and consistent theme throughout all studies. Broad discrimination has also been observed for ML studies in cancer genomics [44]; more established fields with clearer predictor–response relationships, such as medical imaging, are much more consistent [45].

Issues relating to ROB often rest on distinctions in methodology between clinical prediction modelling, ML and genetic association studies. For instance, genetic studies most commonly employ a case–control design. Such studies are extremely useful for identifying genetic risk factors for rare outcomes, but are considered inadequate for prediction modelling as absolute risks cannot be estimated; instead, case–cohort, nested case–control, or prospective cohort designs are preferred [46]. Case–cohort and nested case–control designs involve sampling from an existing cohort and can be used for prediction models if the sampling fraction in controls is accounted for in analysis [47]. To project the prediction to the whole population in case–control studies, positive and negative predictive values should be corrected in accordance with the disease prevalence in the population and ratio of cases and controls in the sample [48]. Similarly, univariable tests of association are applied routinely in GWAS, and are often used in selection of predictors for genetic prediction models. Their application in prediction modelling though is usually discouraged, as predictors may differ in their importance when evaluated in isolation as compared with when considered concurrently with other variables [49].

Lack of adherence to appropriate procedures for ML are also a common cause of a model being assessed as at high ROB. Standard model validation procedures were followed by some researchers; however, many ‘leaked’ information between training and testing sets through not applying predictor manipulations or selection in only the training set/fold, or using the testing set/fold to adjust model hyperparameters, which can impose significant bias on estimates of prediction performance [50].

Most studies provided a measure of classification or discrimination for each model; none reported a measure of calibration. Model calibration compares observed and predicted probabilities of the outcome occurring, and is a crucial part of model development [51] which has been noted for its absence in genetic prediction literature [52]. Authors reporting only classification measures, such as accuracy, sensitivity or specificity, should also note that measures of discrimination are preferred as they use all the information over predicted probabilities and delay any thresholding of risks to a more appropriate time. Of discrimination measures, the AUC is the most widely used in both ML and genetics [53, 54].

Hyperparameter optimisation is an essential part of developing ML models as it determines how they navigate the bias-variance trade-off and learn from data [55]. It is therefore surprising that it was so often unreported or subject to a small number of manual experiments. Hyperparameters should be systematically searched to ensure a model is not over or under-fit. Randomised search has been shown to be more effective than grid search where two or more such parameters require tuning [56], though grid search is also recommended by practitioners for SVMs, often with an initial ‘coarse’ search followed by a more thorough exploration of a finer grid of values [57]. The importance of search is particularly relevant in domains where there are a small number of events per candidate predictor [58], such as genomics, as appropriate hyperparameter choices can reduce overfitting.

Split-sample approaches were used by several studies, but should be avoided in favour of resampling methods such as bootstrapping or *k*-fold CV [59]. The latter is an appropriate form of internal validation for traditional statistical methods; however, estimated prediction accuracies become overly optimistic if done repeatedly, as when used for hyperparameter tuning through repeated rounds of CV. Nested CV, where hyperparameters are optimised in an inner-fold and evaluated in the outer-fold, has been shown to give more realistic estimates [50, 60] but was not used in any studies. A single study presented both internal and external validation of models [32], for which a large drop in performance is seen upon replication. Though partly due to sample overlap between the development set and the summary statistics used for generating a PRS, difficulty with replication is a wider issue in polygenic risk prediction. Risk scores for psychiatric disorders typically explain a small proportion of variance in a trait [61], with generalisation issues compounded by variants with small effect sizes and different allele frequencies between populations. Risk scores generated through ML methods have the potential to be more affected by these issues if appropriate modelling procedures are not followed.

A source of bias not explicitly covered in PROBAST is population structure. Genetic ancestry has the potential to bias both associations [62, 63] and predictions [64, 65] from genetic data. Supervised ML methods have proved particularly sensitive in detecting ancestry [66–68]. Few researchers discussed visualising ancestry or reported exclusions, and none reported modelling adjustments, even when previous association studies on the same datasets had demonstrated stratification and included principal components as covariates. The extent of the bias introduced in these studies is not clear: evidence mostly relates to deliberately predicting populations in humans using ML or looking at bias in complex trait prediction from PRS. While the potential for population stratification to impact predictions is apparent, the method for dealing with it when using ML methods is not. Several techniques have been proposed, including modifications to random forests [69]; exclusions by, or inclusion of, principal components; and regressing-off the linear effects of principal components on SNPs before modelling (for example [70, 71]). Whether any combination of these is sufficient to reduce the effects of population stratification in non-linear ML predictions has not been demonstrated.

General reporting guidelines for ML prediction models are yet to be developed [72], though recommendations for undertaking [73, 74] evaluating [75] or reporting [76] exist for ML in omics data, psychiatry and medicine, respectively, in addition to reporting guidelines outside of ML [21, 77]. We encourage authors to report on implementation, samples, predictors, missingness, hyperparameters and handling of potential information leakage, and consult guidelines where needed. Finally, we advocate for ML methods to be reported alongside PRS as a standard baseline model for comparison. The potential for ML methods to provide improved prediction has received heightened attention in recent years. Any such outcome cannot occur without adherence to standards for the development, validation and reporting of models.

Acknowledgements The authors wish to thank the Dementia Research Institute (UKDRI-3003) and MRC Centre for Neuropsychiatric Genetics and Genomics Centre (MR/L010305/1) and Program Grants (MR/P005748/1).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. J Mach Learn Res. 2011;15:315–23.
2. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag. 2012;29:82–97.
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
4. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst. 2014;27:3104–12.
5. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392–404.
6. Krystal JH, Murray JD, Chekroud AM, Corlett PR, Yang G, Wang X-J, et al. Computational psychiatry and the challenge of Schizophrenia. Schizophr Bull. 2017;43:473–5.
7. Schnack HG. Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). Schizophr Res. 2019;214:34–42.
8. Tandon N, Tandon R. Will machine learning enable us to finally cut the gordian knot of Schizophrenia. Schizophr Bull. 2018;44:939–41.
9. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
10. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99:323–9.
11. Okser S, Pahikkala T, Aittokallio T. Genetic variants and their interactions in disease risk prediction—machine learning and network perspectives. BioData Min. 2013;6:5.
12. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 2014;10:e1004754.
13. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016;46:2455–65.
14. Librenza-Garcia D, Kotzian BJ, Yang J, Mwangi B, Cao B, Pereira Lima LN, et al. The impact of machine learning techniques in the study of bipolar disorder: a systematic review. Neurosci Biobehav Rev. 2017;80:538–54.
15. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord. 2018;241:519–32.
16. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019;24:1583–98.
17. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet 2019;10:267.
18. Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, Duncan L, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360:eaap8757.
19. Kapur S, Phillips A, Insel T. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry. 2012;17:1174–9.
20. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744.
21. Janssens ACJ, Ioannidis JP, van Duijn CM, Little J, Khoury MJ. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Genome Med. 2011;3:16.
22. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.

23. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019; 170:51.
24. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6:e1000097.
25. Pirooznia M, Seifuddin F, Judy J, Mahon PB, Potash JB, Zandi PP, et al. Data mining approaches for genome-wide association of mood disorders. Psychiatr Genet. 2012;22:55–61.
26. Guo Y, Wei Z, Keating BJ, Hakonarson H, The Genetic Consortium for Anorexia Nervosa, The Wellcome Trust Case Control Consortium 3, et al. Machine learning derived risk prediction of anorexia nervosa. BMC Med Genomics. 2016;9:4.
27. Vivian-Griffiths T, Baker E, Schmidt KM, Bracher-Smith M, Walters J, Artemiou A, et al. Predictive modeling of schizophrenia from genomic data: comparison of polygenic risk score with kernel support vector machines approach. Am J Med Genet Part B Neuropsychiatr Genet. 2019;180:80–5.
28. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35: 34–41.
29. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.
30. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Hum Genet. 2014;133:639–50.
31. Acikel C, Son YA, Celik C, Gul H. Evaluation of potential novel variations and their interactions related to bipolar disorders: analysis of genome-wide association study data. Neuropsychiatr Dis Treat. 2016;12:2997–3004.
32. Chen J, Wu J, Mize T, Shui D, Chen X. Prediction of Schizophrenia diagnosis by integration of genetically correlated conditions and traits. J Neuroimmune Pharmacol. 2018;13:532–40.
33. Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A. Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. Am J Med Genet Part B Neuropsychiatr Genet. 2019;180:103–12.
34. Aguiar-Pulido V, Seoane JA, Rabuñal JR, Dorado J, Pazos A, Munteanu CR. Machine learning techniques for single nucleotide polymorphism—disease classification models in schizophrenia. Molecules. 2010;15:4875–89.
35. Yang H, Liu J, Sui J, Pearlson G, Calhoun VD. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of Schizophrenia. Front Hum Neurosci. 2010;4:192.
36. Aguiar-Pulido V, Gestal M, Fernandez-Lozano C, Rivero D, Munteanu CR. Applied computational techniques on Schizophrenia using genetic mutations. Curr Top Med Chem. 2013;13: 675–84.
37. Engchuan W, Dhindsa K, Lionel AC, Scherer SW, Chan JH, Merico D. Performance of case-control rare copy number variation annotation in classification of autism. BMC Med Genomics. 2015;8:S7.
38. Laksshman S, Bhat RR, Viswanath V, Li X, Sundaram L, Bhat RR, et al. DeepBipolar: identifying genomic mutations for bipolar disorder via deep learning. Hum Mutat. 2017;38:1217–24.
39. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.
40. Ghafouri-Fard S, Taheri M, Omrani MD, Daaee A, Mohammad-Rahimi H, Kazazi H. Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. J Mol Neurosci. 2019;68:515–21.
41. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506:185–90.
42. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
43. Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat 2017;38:1182–92.
44. Patil S, Habib Awan K, Arakeri G, Jayampath Seneviratne C, Muddur N, Malik S, et al. Machine learning and its potential applications to the genomic study of head and neck cancer—a systematic review. J Oral Pathol Med. 2019;48:773–9.
45. Islam MM, Yang HC, Poly TN, Jian WS, Li YCJ. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis. Comput Methods Prog Biomed. 2020;191:105320.
46. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683–90.
47. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KGM. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:1–7.
48. Kallner A. Bayes' theorem, the roc diagram and reference values: definition and use in clinical diagnosis. Biochem Med. 2018;28:16–25.
49. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49:907–16.
50. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14:e0224365.
51. Steyerberg EW. Clinical prediction models. 2nd ed. Springer Nature, Switzerland; 2019.
52. Janssens ACJ, Ioannidis JP, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. Eur J Hum Genet. 2011;19:615.
53. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30:1145–59.
54. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010;6:e1000864.
55. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York, NY: Springer New York; 2013.
56. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281–305.
57. Ben-Hur A, Weston JA. User's guide to support vector machines. In: Data mining techniques for the life sciences. Humana Press, New York, NY; 2010. p. 223–39.
58. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ 2015;351:h3868.
59. Steyerberg EW, Harrell FE, Borsboom GJJ, Eijkemans MJ, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
60. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinforma. 2006;7:91.
61. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013;45:984–94.

62. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004;36:512–7.
63. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
64. Belgard TG, Jankovic I, Lowe JK, Geschwind DH. Population structure confounds autism genetic classifier. Mol Psychiatry. 2014;19:405–7.
65. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet. 2017; 100:635–49.
66. Bridges M, Heron EA, O'Dushlaine C, Segurado R, Morris D, Corvin A, et al. Genetic classification of populations using supervised learning. PLoS One. 2011;6:e14802.
67. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 2018;34:301–12.
68. Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol. 2019;36:220–38.
69. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. Nat Commun. 2015;6:7432.
70. Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, et al. Correction for population stratification in random forest analysis. Int J Epidemiol. 2012;41:1798–806.
71. Zheutlin AB, Chekroud AM, Polimanti R, Gelernter J, Sabb FW, Bilder RM, et al. Multivariate pattern analysis of genotype–phenotype relationships in Schizophrenia. Schizophr Bull. 2018;44:1045–52.
72. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577–9.
73. Boulesteix A-L, Wright MN, Hoffmann S, König IR. Statistical learning approaches in the genetic epidemiology of complex diseases. Hum Genet. 2020;139:73–84.
74. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. Nat Mater. 2019;18:422–7.
75. Tandon N, Tandon R. Machine learning in psychiatry—standards and guidelines. Asian J Psychiatr. 2019;44:A1–4.
76. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18:e323.
77. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55.