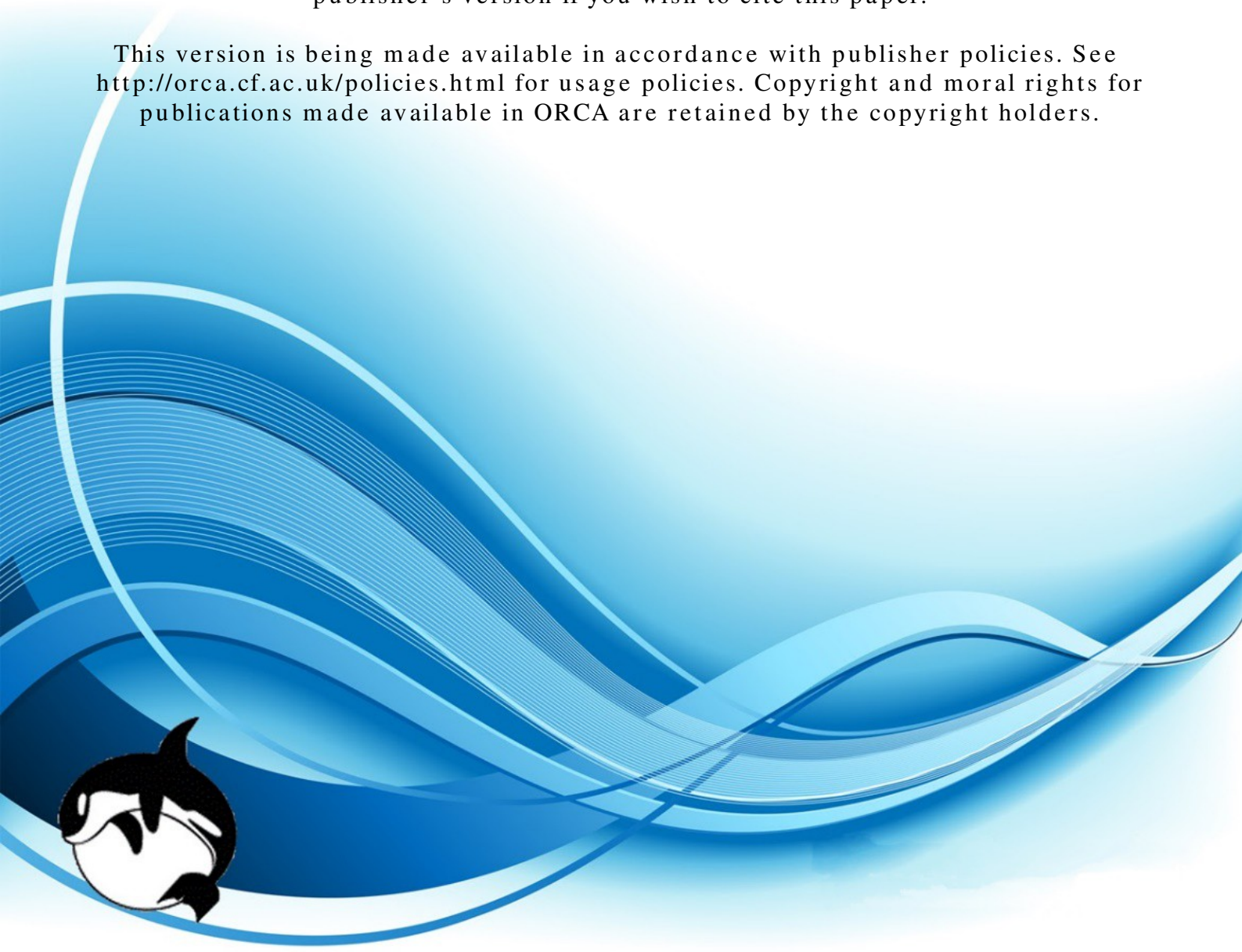

Table of Contents

Supplementary Methods

Inclusion and Exclusion Criteria

Studies which combined genetics and non-genetic data (such as from neuroimaging) were included if they developed or validated a model using only genetic data. Models were only considered for inclusion if they contained two or more genetic predictors from more than one locus. Psychiatric disorders were limited to those with demonstrated heritability and for which large association studies have been undertaken; neurological conditions with psychiatric comorbidities were excluded. A machine learning or statistical learning method was required to be used as the prediction model, with models only using ML for quality control or predictor selection not considered. Changes were made to the registered protocol (registration number CRD42019128820) to further restrict the review's scope, and to clarify inclusion and search criteria before completing database searches.

Extraction

Events per candidate predictor were extracted for all models. Candidate predictors include all predictors considered for inclusion in a model by their association with the outcome. Predictors removed due to association only with other predictors were not counted. As coding of variables is not supplied by most authors, categorical predictors that may be converted to multiple indicator variables by methods are considered only as a single candidate predictor. Similarly, where methods consider additional parameters in the model, such as hidden layers in deep neural networks, only the number of actual predictors is used, not including all possible additional parameters estimated in the model. EPV should

therefore be considered an upper bound. Where authors were ambiguous in their reporting of sample size or number of predictors, bounds of the highest and lowest possible EPV are given.

Where models were fit and internally evaluated before external validation in a single study, we extracted information for both internal and external validation. Internal validation is taken to be any form of evaluation on a subset of the same sample used for training, including splitting samples between training and test sets, bootstrapping and *k*-fold cross validation. Apparent validation, where training and testing are both done on the whole sample, is also recorded under internal validation for the purpose of this review. This is part of model development. External validation is understood as evaluation on an independent dataset, which differs in temporal, geographic or other aspects, and is not simply a splitting-off from the original sample. If multiple models were presented with subsampled predictors or participants, only main models presented in the text were extracted; if such a distinction was unclear, all models were selected for review. Where AUC was only available graphically it was extracted from the figure using Plot Digitizer [1], and accuracy was calculated from the confusion matrix if not provided in-text.

Model discrimination was extracted independently by two authors (MBS, KC). AUC extracted were the same for both authors, except for 3 of 77 models from a single study; consensus was reached after reviewing the text. Studies often included many models; logistic regression models were only extracted where they received the same predictors as ML methods, in order to keep models comparable.

PROBAST

Risk of bias (ROB) was assessed using the prediction model risk of bias assessment tool (PROBAST). Where information was unavailable within a study, any references or links given to descriptions of datasets or methods were examined. Questions remained unchanged; however, recommendations for assessing studies using genetics and machine learning were added to adapt the tool and keep consistency in answers across models and reviewers. These are detailed below. No studies dictated if a model was intended for prognostic or diagnostic use. For the purpose of assessing ROB, models are assumed to be diagnostic; changing intended model use to prognostic does not alter the final ROB assessments for models. Where databases or publications were referenced for a study, these were assessed for information relevant to ROB. As large genetic datasets may change composition over iterations as smaller studies are added, additional publications that may describe an iteration of a publicly available dataset, but which were not referenced in the included study, were not examined.

Questions in PROBAST are formatted such that answering “Yes” indicates low risk of bias, and answered “no” indicates high risk of bias. Normally, if any questions within a domain are rated “no” or “probably no” (N/PN), then the rating is considered to be “high” ROB for that domain. In the absence of any N/PN responses, if any questions are reported as “no information” (NI), then the domain is taken to have “unclear” ROB. If instead all questions were answered as “yes” or “probably yes” (Y/PY), then the domain is rated as “low” ROB. Select situations where questions are rated NI or N/PN were allowed to be rated “low” ROB overall. For predictors, if question 2.2 (“Were predictor assessments made without knowledge of outcome data?”) was rated as NI or Y/PY, overall rating for ROB of predictors

was allowed to be “low”. Knowledge of the outcome can enable careful design of cases and controls across arrays and batches, and exclusion by a more stringent threshold of Hardy-Weinberg equilibrium in controls. These may allow for reduced ROB for predictors, rather than increased. For outcome, question 3.5 (“Was the outcome determined without knowledge of predictor information?”), if NI or Y/PY, was allowed to be rated “low” ROB for outcome overall if it was considered that genotypes or other predictors would have been extremely unlikely to influence the outcome of standard assessments, or that outcomes were likely to have been assessed prior to genotyping. For question 4.1 (“Were there a reasonable number of participants with the outcome?”), events per candidate predictors were assessed against recommendations using machine learning methods with default hyperparameters, and therefore represent the wort-case scenario. If EPV was determined to be near to the cut-off, and all other modelling procedures indicated low ROB, including appropriate regularisation and handling of predictors, analysis was allowed to be rated “low” overall. In practice, this situation did not occur.

PROBAST requires a ROB assessment of each evaluation of each distinct model [2]. Development and validation are therefore both assessed for each model and contribute separately to overall counts. Restricting counts to development-only does not appreciably change results. ROB was assessed for all studies by one author (MBS), with the exception of a single publication on which MBS and VEP are co-authors [3]. Here two authors, MBS and KC, independently assessed ROB, the latter being uninvolved in the original study. Differences were overcome through consensus. A third colleague not included in the original study was designated as arbiter should disagreements be unable to be resolved. This situation did not occur.

1.1 Were appropriate data sources used, e.g. cohort, RCT, or nested case-control study data?

Studies may be made of multiple smaller studies, some of which are cohorts or where cases are from cohorts but controls are from elsewhere. If cases and controls are sampled from different sources to give a roughly balanced (equal events and non-events) combined sample, denote the combined sample as case-control. If absolute risk cannot be estimated from the combined sample, rate as N/PN.

1.2 Were all inclusions and exclusions of participants appropriate?

If the target population for the prediction model is undefined, rate as NI, as this cannot be assessed.

2.1 Were predictors defined and assessed in a similar way for all participants?

If genotypes measured on different arrays and there has been no effort to demonstrate similarity across arrays or lack of batch effects, rate N/PN. If genotypes from different arrays have been imputed to the same panel of reference genomes to infer untyped or missing variants, rate Y/PY.

3.1 Was the outcome determined appropriately?

Consensus best-estimate diagnosis using medical records and structured interview is considered appropriate. Use of only a structured interview is also considered appropriate, but use of only interviews with family members and records is rated N/PN. Routine care registry data are appropriate only if studies confirming comparability with standard

diagnostic methods are available. If method is appropriate only for cases, rate 3.1 as Y and 3.4 as N.

3.2 Was a prespecified or standard outcome definition used?

Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Classification of Diseases (ICD)-based outcomes are accepted.

3.4 Was the outcome defined and determined in a similar way for all participants?

If the same assessments tool was used for all participants, rate Y/PY. If cases were assessed differently to controls, rate N/PN.

3.6 Was the time interval between predictor assessment and outcome determination appropriate?

If predictors are genetics-only, rate Y/PY. If predictors include gene-expression data sampled after diagnosis or onset, rate N/PN.

4.1 Were there a reasonable number of participants with the outcome?

No recommendations are available for assessing events per variable (EPV) in machine learning models. To our knowledge, only one paper has attempted to assess EPV needed for machine learning models across multiple datasets [4], which we use here as a guide in lieu of a more rigorous alternative. For the purpose of assessing ROB in this review, support vector machines are required to have greater than 200 EPV. Neural networks require at least 200 EPV, but a cut-off of at least 500 EPV should be imposed as architecture can vary greatly. Random forests are also required to have greater than 500 EPV. For other machine

learning methods not specified above, 200 EPV is taken as the minimum requirement. Everything below these cut-offs is rated as N/PN. It should be noted that the models these estimates are based on were run using default (hyper)parameters [4] on non-genetic data. Final assessment of ROB for "analysis" should therefore take into account regularisation and model architecture, as models with an EPV of less than 200 may still be rated as "low" ROB for the domain. However, given that all models had multiple aspects of analysis which introduced ROB, changing these thresholds would not affect the final rating for the 'analysis' domain in any models.

4.4 Were participants with missing data handled appropriately?

For imputation using a genetics-specific application or server, such as IMPUTE2, rate Y/PY. For imputation in the sample using other methods, rate N/PN. For complete-case analysis, rate N/PN.

4.5 Was selection of predictors based on univariable analysis avoided?

If any plink-based univariable tests for association in the current dataset were used, rate N/PN. If information from an external published GWAS was used to select predictors, rate Y/PY.

4.8 Were model overfitting, underfitting, and optimism in model performance accounted for?

If nested cross-validation was used, rate Y/PY, assuming other standard procedures were followed. If any method of repeated cross-validation on the whole dataset where both tuning and evaluation of models were done in the same *k*-fold cross-validation loop was used, or where test data were observed during tuning of hyperparameters, rate N/PN.

4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?

If no model coefficients or assigned weights clearly reported, rate NI, as this cannot be assessed.

Supplementary Figures

Figure S1: PRISMA flow diagram. Where a publication met multiple exclusion criteria, it is counted only under the first reason in the list.

Figure S2: within-study risk of bias and applicability assessed by PROBAST. Colours indicate low, high or unclear risk of bias or applicability.

Figure S3: discrimination (AUC) for machine learning, logistic regression and polygenic risk scores. Internal validation (split-sample) and partly-external validation (with sample overlap) are reported for the same models in a single study [5]. [1]Median AUC for internal validation (model development). [2]Median AUC for external validation (independent replication). Annotated scores are the median AUC for each model and study. Pirooznia et al. (bipolar disorder) and Vivian-Griffiths et al. (schizophrenia) show SNP-only models for LR and ML [3, 6], while Chen et al. (schizophrenia) used multiple schizophrenia-associated trait polygenic risk scores as predictors [5]. PRS model performance was extracted from a figure when unreported in-text [3]. AUC is shown only for 5 of the 9 reported logistic regression models; a fourth study compared ML and LR but did not report discrimination [7]. AUC was not available for a logistic regression which was reported as attempted but not completed for one study [6]. AUC: area under the receiver operating characteristic curve, ML: machine learning, LR: logistic regression, PRS: polygenic risk scores.

Supplementary Tables

Where percentages are reported in any table, they are taken from the total number of models, 77, and rounded to the nearest integer unless stated otherwise. Some aspects of methodology differed between models within studies. Where this occurs, studies are counted under each category that has been met unless stated otherwise, and total counts may not sum to 13.

Search

1. (schizophreni* or schizoaffective or schizotyp* or anxiety or depressi* or autis* or adhd or anorexi* or bullimi* or psychos?s or psychotic or manic or mania or hypomani* or tourette* or obsessive compulsive disorder or ocd).ti,ab. or (exp SCHIZOPHRENIA/ or Bipolar Disorder/ or exp ANXIETY DISORDERS/ or exp Autism Spectrum Disorder/ or exp Depressive Disorder/ or Attention Deficit Disorder with Hyperactivity/ or Anorexia Nervosa/ or Bulimia Nervosa/ or exp Obsessive-Compulsive Disorder/ or Tourette Syndrome/)
2. (machine learning or statistical learning or pattern analysis or pattern recognition or ensemble or bayesian network* or relevance vector machine* or support vector machine* or decision tree* or classification tree* or regression tree* or elastic net or bagging or gradient boosting or neural network or perceptron or nearest neighbo?r or gaussian process* or ridge or lasso or regulari#ed regression or penali#ed regression or naive bayes or (deep adj3 learning) or (boosted adj2 trees) or (deep adj2 network) or (random adj2 forest) or (supervised adj2 learning)).ti,ab. or exp Machine Learning/
3. (rare variant* or rare variation or copy number variant* or copy number variation* or dna variant* or polygenic or genetic* or polymorphism* or genotype* or genome* or genomic* or exome*).ti,ab. or exp Polymorphism, Genetic/

4. 1 and 2 and 3
5. limit 4 to english language
6. limit 5 to journal article
7. remove duplicates from 6

Table S1: example literature search from Medline (Ovid).

Extraction

Domain	Item
Background	Reference
	Disorder
	Study design
	Publication number
	Model type (diagnostic/prognostic)
Participants	Recruitment method
	Study setting
	Retrospective or Prospective?
	Number of Centres
	Inclusion/Exclusion criteria
	Sample description
	Study Dates
	Dataset names or identifiers
Sample size	Total number of observations before QC
	Total number of observations after QC
	Case:control ratio in final dataset
	Number of cases in training set/fold
	Events Per Variable in the training set/fold

Outcome	Definition of outcome
	Measurement
	Same for all patients?
	Type of outcome (single/combined)
	Were assessors blinded to knowledge of predictors?
	Predictors in outcome?
Predictors	Genotyping/sequencing method
	Imputation method and reference
	Types of genetic data
	Method of choice of variants to genotype/sequence
	Genetic Predictor QC
	Number of candidate predictors
	Number predictors in final model
	Coding of genetic data
	Risk allele definition for coding at a single locus
	Knowledge/annotation information included?
	Knowledge/annotation inclusion method
	Was measurement of predictors blinded to outcome/other predictors?
	Any other handling of predictors
	Was leakage handled appropriately?
Participant QC	Genetic sample QC
	Method for accounting for genetic ancestry
	Method of accounting for plate/batch/site effects
	Method for accounting for relatedness
Missing Data	Number participants with any missing value[1,2]
	Number of participants with missing data for each predictor[1]
	Handling of missing data
	Modelling method/representation

Model	Model implementation (programming language)
Development	Model modifications
	Predictor selection types used
	Method for selection of predictors prior to modelling (filter)
	Method for selection of predictors during modelling (wrapper)
	Method for selection of predictors as part of model (embedded)
	Hyperparameter search method
	Tuned Hyperparameters
	Class imbalance method[1]
Model	Discrimination measures reported
Performance	Calibration measures reported
	Classification measures reported
	Other measures reported
	A-priori decision threshold cut-off used for classification?
Model Evaluation	Method for testing model performance internally
	Method for testing model performance externally
	Model adjusted or updated after poor validation?[3]
Results	Model AUC
	Model Accuracy, sensitivity and specificity
	Model calibration[1]
	Comparison of distribution of predictors[1]
	Data/code available (link)
Extra	Resources
	Notes

Table S2: extraction form, modified from the checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist [8]. Items which overlap heavily with prediction model risk of bias assessment tool (PROBAST) signalling questions, such as participant information, are reported in risk of bias summaries.

AUC: area under the receiver operating characteristic curve, QC: quality control. [1]Not reported in any publications.

[2]Number of participants excluded above a threshold of missingness was reported in many studies. [3]No for all publications.

Samples

Datasets

Titles and descriptions of studies making up a dataset are recorded as given in the extracted publication. Where references are supplied, these were given in the text, or clear from an online repository, such as the database of Genotypes and Phenotypes (dbGaP) [9]. Where datasets appear to overlap, this has been noted.

Study	**Disorder**	**Dataset**
Yang et al. (2010)	Schizophrenia	No name/reference given
Ghafouri-Fard et al. (2010)	Autism	No name/reference given
Aguiar-Pulido et al. (2010;2013)	Schizophrenia	External sample[a]
Wang et al. (2018)	Schizophrenia	PsychENCODE[b]
	Bipolar disorder	PsychENCODE[b]
	Autism	PsychENCODE[b]
Pirooznia et al. (2012)	Bipolar disorder	BGSC[c†‡] (DEV), WTCCC[d*] (VAL)
Laksshman et al. (2017)	Bipolar disorder	Not clearly reported[e]
Acikel et al. (2016)	Bipolar disorder	Whole-Genome Association Study of Bipolar Disorder[f†‡]
Li et al. (2014)	Bipolar disorder	Whole-Genome Association Study of Bipolar Disorder[f†‡]
	Schizophrenia	Genome-Wide Association Study of Schizophrenia[g†]

Guo et al. (2016)	Anorexia	GCAN[h], WTCCC[d*], CHOP[i], PFCG[j]
Trakadis et al. (2019)	Schizophrenia	Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing[k**]
Engchuan et al. (2015)	Autism	AGP[l]
Chen et al. (2018)	Schizophrenia	MGS[m†], SSCCS[n**] (DEV), CATIE[o†] (VAL)
Vivian-Griffiths et al. (2019)	Schizophrenia	CLOZUK[p*]

Table S3: sample overlap between studies. **a**: Galician sample described elsewhere [10]. **b**: PsychENCODE, made up of 8/9 studies, where only 6 are listed in the supplementary as having genotype data - study 1 (BrainGVEX, consisting of the Banner Sun Mental Research Institute, BSHRI [11], and Stanley Medical Research Institute, SMRI); study 2 (BrainSpan), no genotype data; study 3 (CommonMind [12]); study 4 (Yale-ASD); no genotype data; study 5 (UCLA-ASD [13]); study 6 (BipSeq); study 7 (CMC_HBCC); study 8 (LIBD_szControl + BipSeq); study 9 (not reported). Information and data also available through an online repository [14]. **c**: Bipolar Genome Studies Consortium (BGSC) [15], made up of the Genetic Association Information Network European American (GAIN) [16], and the Translational Genomics Research Institute (TGRI) samples. Controls obtained through Knowledge Networks (KN) [17], and recruitment described elsewhere [18, 19]. **d**: Wellcome Trust Case Control Consortium (WTCCC). Bipolar Disorder cases are described in methods, with further information provided elsewhere [20, 21]. Controls include the 1958 British Birth Cohort (58BC) [22] and the UK Blood Service (UKBS) [23]. **e**: part of the Critical Assessment of Genome Interpretation (CAGI)-4 challenge. Laksshman et al. [24] reference Daneshjou et al. [25], from which a third reference [26] gives information on an exome dataset with only bipolar cases recruited for a suicide study, but not controls. **f**: Whole-Genome Association Study of Bipolar Disorder, dbGaP study accession "phs000017.v3.p1". References on dbGaP provide further details on sample recruitment [18, 27]. Acikel et al. acquired Bipolar Disorder Only (BDO) participants [28]; Li et al. report using the Bipolar and Related Disorders (BARD) subset [29]. Controls, obtained through KN, are described under "Clinical Procedures" of the relevant dbGaP entry, and by other studies [17]. **g**: Genome-Wide Association Study of Schizophrenia, dbGaP study accession "phs000021.v3.p2". Cases described on dbGaP, controls obtained through KN. **h**: the Genetic Consortium for Anorexia Nervosa (GCAN). **i**: Price Foundation Collaborative Group and the Children's Hospital of Philadelphia (CHOP). Methodological details for Guo et al. are also referenced to a previous study [30]. **j**: the Price Foundation Collaborative Group (PFCG). **k**: Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing, dbGaP study accession "phs000473.v1.p1". Described in more detail

elsewhere [31]. **l**: Autism Genome Project (AGP); three references supplied for methodology and participants [32–34]. **m**: Molecular Genetics of Schizophrenia (MGS) [35], with controls from KN. **n**: Swedish Schizophrenia Case Control Study (SSCCS) [36]. **o**: Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) [37, 38], with controls from KN. Imputation for Chen et al. is also given elsewhere [39]. **p**: CLOZUK [40]; controls from 58BC and UKBS. *Includes controls from the 1958 British Birth Cohort and the UK Blood Service. †Includes controls from Knowledge Networks. ‡Publications do not all give the same dataset name or description, but do include a common reference for recruitment or inclusion criteria. **Studies refer to a Swedish population-based sample with the same outcome definition, but no clear statement or reference describing sample overlap.

Handling of Missing Data

Method	**Studies**	**Models**
Reported	7	43 (56%)
Exclusion (complete-case analysis)	1	1
Code missingness as category in predictor	1	12
Imputation after excluding high missingness	5	30
Imputation using genetics server/application	3	16
Imputation in-sample from binomial distribution[a]	2	14
Unclear/unreported	6	34 (44%)
Only exclusion for high missingness reported[b]	4	28
Not reported	2	6

Table S4: missingness. Handling of missing data differed between the development and validation set for Pirooznia et al. (2012), where imputation is only reported for external validation [6]; these models are counted under the method reported in model development, "only exclusion for high missingness". [a]A study [3] reported using unspecified imputation prior to quality control filters, before a second in-sample imputation and is recorded once as in-sample. [b]Includes high missingness filters for samples, predictors or both, with method for handling remaining missingness not reported.

Software

Language/Implementation/Method	Studies	Models
R	4	11 (14%)
glmnet (LASSO)	1	1
randomForest (RF)	2	2
party (CIF)	1	1
e1071 (SVM, NB)	2	2
gbm (GBM)	1	1
XGBoost (Histogram-based GBM)	1	1
kNN (*k*-NN)	1	1
MDR (MDR)	1	2
Python	4	16 (21%)
scikit-learn	3	12
SVM	1	8
Data handling	1	1
Unspecified	1	3
Keras (NN)[a]	2	4
Tensorflow (NN)	1	4
Java (WEKA)[b]	2	28 (36%)
Matlab	2	11 (14%)
Matlab (NN)	2	10
libSVM (SVM)	1	1
Not reported	3	11

Table S5: software and packages used in machine learning. [a]Backend to Keras not specified. [b]Methods used in WEKA: neural networks (linear, perceptron and radial basis function), evolutionary computation, multifactor dimensionality

reduction, Bayesian networks, naïve Bayes, support vector machine, decision tables, decision tree-naïve Bayes, best-first tree, AdaBoost. LASSO: least absolute shrinkage and selection operator, RF: random forest, CIF: conditional inference forest, GBM: gradient boosting machine, XGBoost, eXtreme Gradient Boosting, *k*-NN: *k*-nearest neighbours, MDR: multifactor dimensionality reduction, SVM: support vector machine, NN: neural network, NB: naïve Bayes.

Bias

Method of accounting for ancestry

Method	Studies	Models
Population substructure identified in current study but not accounted for	2	14 (18%)
Visualised by PCs for subsample after restricting to European	1	9
Table of ancestry for European American and African American[a]	1	5
Unclear[b]	9	50 (65%)
Population structure identified in dataset reference(s)	7	42
Exclusion of non-European ancestry through PCs/MDS	5	35
Visualised but observations not excluded	3	11
Reported as European/Caucasian-only, no details given	2	8
Not reported in publication or reference	2	13 (17%)

Table S6: methodology for accounting for population structure. Where development or validation sets are made-up of multiple datasets with separate ancestry filters, these are counted separately. [a]Method of establishing ancestry not specified. [b]Ancestry not clearly specified in current study. PCs: principal components, MDS: multi-dimensional scaling.

Models

Model Performance Measures

Reported measures	Studies	Models
Discrimination	8	45 (58%)
AUC	8	45
ROC plot	4	7
Classification	9	41 (53%)
Accuracy	8	39
Sensitivity/Recall/Hit-rate/TPR	6	16
Specificity/TNR	4	10
F_1-score (*F*-measure)	3	12
Precision/PPV	3	12
Confusion matrix	3	3
Other	5	29 (38%)
Variance explained on liability scale	1	9
p-value*	1	4
% correctly classified cases, averaged over repeats	1	4
Nagelkerke's pseudo-R^2	1	4
t-test comparisons between models	1	8

Table S7: model performance. *The *p*-value "indicates that XGBoost algorithm is performing better than a random predictor simply predicting the majority class" [41]. ROC: receiver operating characteristic, AUC: area under the ROC curve, TRP: true positive rate, TNR: true negative rate, PPV: positive predictive value. As many studies reported multiple measures, percentages do not combine to 100.

Decision threshold cut-off

Method for choosing cut-off	Studies	Models
a-priori	1	9 (22%)
Unclear	3	6 (15%)
Unreported	5	26 (63%)

Table S8: method for choosing decision threshold when reporting classification metrics. Studies which were unclear either reported a general outline of how classification works for a given method, without stating this was used in the current implementation, or reported the use of 0.5 as the threshold but not how the number was chosen. Percentages are taken from the total number of models which reported classification measures, 41, and rounded to the nearest integer. Number of studies does not sum to 13 as not all studies reported classification metrics.

Validation

Method	Studies	Models
Internal validation		
Cross-validation	8	44 (57%)
3-fold	1	4
4-fold	1	8
5-fold	2	8
10-fold	3	22
LOOCV	1	2
Split-sample	5	16 (21%)
34% train[a]	1	3
40% train[b]	1	3
70% train	1	4
80% train	1	2

90% train	1	4
Apparent	1	1 (1%)
Not reported[c]	1	16 (21%)
External Validation		
External (temporal, geographic)[c]	1	16 (21%)
Partly external[d]	1	4 (5%)
Not performed	11	57 (74%)

Table S9: validation. Percentages are given with respect to 77, the total number of models. Methodology for internal validation differed between models in a study [31], which is counted in cross-validation (CV), split-sample and apparent. [a]Approximately equal three-way split between predictor selection, train and test, with 10-fold CV performed in the training fold for hyperparameter tuning. [b]40% train, 10% test, 50% final test. [c]No performance measures reported for internal validation, but discrimination for fully external validation reported [25]. [d]Control sample used in development and validation partially overlaps. LOOCV: Leave-one-out cross validation.

Study	Method	Data	Modifications	n/N	p/P	Imbalance	EPV	Risk allele	Sensitivity	Specificity	Validation
a	AB	SNP	Y	20/40	150/367	1	0.0054	NR	0.7175	0.76	CV
a	SVM	SNP	N	20/40	367/367	1	0.0054	NR	0.4	0.4	CV
b	NN	SNP	Y	487/942	15/15	0.93	32.5	NR	0.8275	0.6395	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	EC	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	MDR	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	BN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	SVM	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	DTb	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	DTNB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	BFT	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	AB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
d	NN	SNP/GE	N	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	N	94/188	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	94/188	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	94/188	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	N	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
e	SVM	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext

e	NN	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
f	NN	Exome	Y	200/400	~1000/>500000*	1	<0.0004*	Ref/alt	0.64	NR	Split
f	RF	Exome	N	200/400	~1000/>500000*	1	<0.0004*	Ref/alt	0.55	NR	Split
f	DT	Exome	N	200/400	~1000/>500000*	1	<0.0004*	Ref/alt	0.54	NR	Split
g	RF	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.998	NR	App.
g	NB	SNP	N	483/1414	693/761830	1.93	0.00063	NR	0.734	NR	Split
g	*k*-NN	SNP	N	483/1414	693/761830	1.93	0.00063	NR	0.954	NR	Split
g	MDR	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.664	NR	CV
g	MDR	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.883	NR	CV
h	Ridge	SNP	N	653/1158	298604/298604	0.77	0.0022	NR	NR	NR	CV
h	SVM	SNP	N	653/1158	98604/298604	0.77	0.0022	NR	NR	NR	CV
h	LASSO	SNP	N	653/1158	98604/298604	0.77	0.0022	NR	NR	NR	CV
h	Ridge	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
h	SVM	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
h	LASSO	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
i	LASSO	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	0.11	0.97	Split†
i	SVM	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	NR	NR	Split†
i	GBM	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	NR	NR	Split†
j	LASSO	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.720	0.773	Split
j	SVM	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.708	0.706	Split
j	RF	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.820	0.813	Split
j	GBM	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.849	0.866	Split
k	CIF	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	RF	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	SVM	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	NN	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV

l	NN	PRS	N	5018/10859	19/116	1.16	43.26	NR	NR	NR	Split/ Ext.
l	NN	PRS	N	5018/10859	116/116	1.16	43.26	NR	NR	NR	Split/ Ext.
l	NN	PRS	N	5018/10859	14/29-32*	1.16	156.81-173.03*	NR	NR	NR	Split/ Ext.
l	NN	PRS	N	5018/10859	26/29-32*	1.16	156.81-173.03*	NR	NR	NR	Split/ Ext.
m	SVM	SNP	N	3446/7731	125/125	1.24	27.57	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	125/125	1.13	44.43	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	4998/4998	1.24	0.69	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	4998/4998	1.13	1.11	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	125/125	1.24	27.57	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	125/125	1.13	44.43	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	4998/4998	1.24	0.69	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	4998/4998	1.13	1.11	Ref	NR	NR	CV

Table S10: overview of prediction models. n: number of cases used in model development in final model, N: number of total observations in model development in final model, p: number of predictors in final model, P: number of candidate predictors, EPV: events per candidate variable/predictor, NR: not reported, NCR: not clearly reported, Ref: risk allele coded as reference allele, Alt: coded as alternative allele, SNP: single nucleotide polymorphism, CNV: copy number variant, PRS: polygenic risk score, GE: gene expression, AB: AdaBoost, SVM: support vector machine, NN: neural network, EC: evolutionary computation, MDR: multifactor dimensionality reduction, BN: Bayesian networks, NB: naïve Bayes, DTb: decision tables, DTNB: decision table naïve Bayes, BFT: best-first tree (BFTree), RF: random forest, DT: decision tree, *k*-NN: *k*-nearest neighbours, LASSO: least absolute shrinkage and selection operator, GBM: gradient boosting machine, CIF: conditional inference forests, CV: cross-validation, n/a: not applicable. †Study used a roughly equal 3-way split for predictor selection, training and testing, where 10-fold CV was used in the training fold [42]. Splits were repeated, but reported AUCs in the main text are for only one of the repeats; the study is recorded here as split-sample. *Number reported is unclear; upper and lower bounds, or an approximation given by the authors in the text are used. Where insufficient information is provided to give a reasonable approximation for predictors, NCR or NR is recorded. Imbalance refers to class imbalance, given here as number of controls divided by number of cases in model development. Modification refers to whether a classifier was used "out-of-the-box", N, or was modified in some way, Y. Validation is *k*-fold CV, split-sample (Split), apparent (App.) or external (Ext.). A single study reported internal validation (split-sample) and external validation (but with partial sample overlap) [5]. Studies: a (Yang et al., 2010) [43], b (Ghafouri-Fard et al., 2019) [44], c (Aguiar-Pulido et al.,

2010;2013) [45, 46], d (Wang et al., 2018) [7], e (Pirooznia et al., 2012) [6], f (Laksshman et al., 2017) [24], g (Acikel et al., 2016) [28], h (Li et al., 2014) [29], i (Guo et al., 2016) [42], j (Trakadis et al., 2019) [41], k (Engchuan et al., 2015) [47], l (Chen et al., 2018) [5], m (Vivian-Griffiths et al., 2019) [3].

Predictors

Coding of predictors

Coding	Studies	Models
Reported	6	35 (45%)
Continuous (weighted average of additive SNPs; PRS)	1	4
Counts of genes per gene set (CNV)	1	4
Counts of variants per gene (Exome)	1	4
Additive model (0, 1, 2), missing coded as 3 (SNP)	1	12
Z-transformation of additive model (0, 1, 2; SNP)	1	8
One-hot encoded (SNP)	1	3
Unclear/unreported	7	42 (55%)
Unclear[a]	2	3
Not reported	5	39

Table S11: coding of predictors. [a]Coding implied through description as 'ordinal' or through an abstract description of the type of classifier, but not clear.

Information in predictors

Method	Studies	Models
Additional knowledge used	9	49 (64%)
Predictors	8	43
Array not genome-wide	3	15

Predictors only from brain-expressed genes	1	8
Selection by *p*-value cut-off from external GWAS	1	8
Annotation of gene and variant-type	1	4
Annotation of gene and gene set	1	4
Choice of phenotypes and weights from GWAS for SZ-PRS	1	4
Modelling	1	6
Non-zero matrix weights in cRBM determined from GE data	1	6
Unclear/unreported	6	28 (36%)
Not clear	1	3
Not reported	5	25

Table S12: explicit use of additional knowledge in selecting or weighting of predictors and modelling. Implicit knowledge, such as choice of a linear machine learning method, or additive encoding of genotyping data, are not included. GE: gene expression, cBRM: conditional restricted Boltzmann machine.

Predictor selection

Type	**Studies**	**Models**
Filter	8	48 (62%)
Association test in external dataset, clumping	1	8
Association test in current dataset, clumping	1	8
Association test in current dataset for brain-expressed genes only, clumping	1	8
Association test in split of current dataset, *p*-value cut-off	1	3
Pruning, association test in current dataset, *p*-value cut-off	1	5
Embedded (LASSO/RF/GBM combined)[a]	1	4
Embedded (LASSO) with *p*-value cut-off	1	2

Forward sequential feature selection (FSFS)[b]	1	1
Correlation with outcome or intermediate phenotype	1	9
Embedded	8	20 (26%)
Regression (LASSO)	3	4
Tree-based	7	13
RF (including CIF)	4	8
Boosting (GBM, AdaBoost)	3	3
DT	2	2
Other	2	3
DTb	1	1
DTNB	1	1
Feature-selective AdaBoost[c]	1	1
Unclear[d]	1	3 (4%)
None reported	6	18 (23%)

Table S13: predictor selection technique. [a]Trakadis et al. (2019) report predictors being selected "in combination of" embedded methods, but do not state how such methods were combined [41]. [b]FSFS is a wrapper on an embedded method, used as a filter. [c]Yang et al. (2010) modified AdaBoost to include univariable predictor selection within each iteration before training each weak learner [43]; as the modification is within each iteration it is listed as "embedded" here. This is counted once under feature-selective AdaBoost, and is not counted under 'Boosting'. [d]Laksshman et al. (2017) report using "L1-based feature selection" but no indication about what method the L1-norm was applied to [24]. LASSO: least absolute shrinkage and selection operator, RF: random forest, GBM: gradient-boosting machine, DTNB: decision table-naïve Bayes, DTb: decision table, DT: decision tree, CIF: conditional inference forest. Several models exploited both filter and embedded methods; these are counted in both sections.

Leakage handled appropriately?	Studies	Models
Yes/Probably Yes	7	44 (57%)
No/Probably No	7	32 (42%)
Predictor selection performed prior to cross-validation	2	7
Predictor transformed prior to cross-validation[a]	4	22
Prior knowledge in predictors generated from test set	1	4
DEV and VAL sets overlap	1	4
HP chosen by test-set/split performance	4	22
GRN from whole dataset used to set NN architecture	1	6
Unclear[b]	1	1 (1%)

Table S14: handling of information "leaks" during training. Where studies have multiple reasons for suspected leakage, each of these is counted separately. If predictors were reduced to a set number before cross-validation was described, or a transformation was not reported as having been done within a pipeline or for each fold of cross-validation, this is recorded as 'probably no'. [a]Transform includes anything that summarises information from the test set, such the mean of the whole sample in a z-transformation. [b]Predictor handling implied, as scikit-learn is listed for pre-processing and preparation, but no pre-processing steps are given [44]. DEV: development, VAL: validation, HP: hyperparameter, GRN: gene regulatory network, NN: neural network.

Hyperparameter search

Search method for hyperparameters	Studies	Models
Search method reported	4	15 (19%)
Grid	1	1
Random	1	8
Manual	2	12
Bias variance decomposition	1	2
Default hyperparameters	1	16 (21%)

Search method unclear/unreported	9	46 (60%)
Not clearly reported[a]	2	8
Not reported	7	38

Table S15: hyperparameter search technique. [a]Methods reported clearly for other models in publications, but not made clear that the same methods apply to extracted models. One publication [3] used both manual and random elements for search, and is counted in both categories. Manual tuning by Chen et al. (2019) is implied through reported values which were attempted for hyperparameters, but not explicitly stated [5]. Hyperparameters searched systematically using a given set of values are denoted as grid search. If authors report attempting various hyperparameter choices but give no indication of systematic search or value choices, this is recorded as manual. Two studies (12 models) reported hyperparameters that were tuned but gave no indication of how this was done [7, 24]. A study (1 model) reported search methodology, but not what hyperparameters were tuned [43].

Method	**Studies**	**Models**
Reported	6	26 (34%)
SVM (RBF)		
C	2	9
Gamma	2	9
AdaBoost[a]		
Iterations	1	1
Neural Networks		
Epochs	2	12
Optimiser	1	4
Activation function	1	4
Layers	1	4
LASSO		
Lambda	1	1

Unclear/unreported	9	51 (66%)
Not clearly reported	2	5
Not reported	8	46

Table S16: hyperparameters tuned during model training. [a]Feature-selective AdaBoost [43]. Manual experiments with different hyperparameters are presented by Engchuan et al. (2015) in the supplementary: these are included as "not reported", as they appear to be post-hoc experiments rather than a search as part of learning [47]. Several studies report either hyperparameter search method, or the hyperparameters that were tuned, but not both (see Table S15). A study (16 models) used the default hyperparameters (Table S15) and is counted here under 'not reported' [6].

References

1. Huwaldt JA, Steinhorst S. Plot digitizer. 2005.
2. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019; **170**: W1.
3. 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**: 80–85.
4. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014; **14**: 137.
5. 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**: 532–540.
6. 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.
7. 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 (80-). 2018; **362**: eaat8464.
8. 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

421 Modelling Studies: The CHARMS Checklist. PLoS Med. 2014; **11**: e1001744.
422 9. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI
423 dbGaP database of genotypes and phenotypes. Nat Genet. 2007; **39**: 1181–1186.
424 10. Domínguez E, Loza MI, Padín F, Gesteira A, Paz E, Páramo M, et al. Extensive linkage
425 disequilibrium mapping at HTR2A and DRD3 for schizophrenia susceptibility genes in
426 the Galician population. Schizophr Res. 2007; **90**: 123–129.
427 11. Beach TG, Sue LI, Walker DG, Roher AE, Lue L, Vedders L, et al. The Sun Health
428 Research Institute Brain Donation Program: description and experience, 1987-2007.
429 Cell Tissue Bank. 2008; **9**: 229–245.
430 12. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. Gene
431 expression elucidates functional impact of polygenic risk for schizophrenia. Nat
432 Neurosci. 2016; **19**: 1442–1453.
433 13. Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, et al.
434 Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in
435 autism. Nature. 2016; **540**: 423–427.
436 14. PsychENCODE Integrative Analysis.
437 https://www.synapse.org/#!Synapse:syn4921369/wiki/235539. Accessed 28
438 November 2019.
439 15. Mahon PB, Payne JL, MacKinnon DF, Mondimore FM, Goes FS, Schweizer B, et al.
440 Genome-Wide Linkage and Follow-Up Association Study of Postpartum Mood
441 Symptoms. Am J Psychiatry. 2009; **166**: 1229–1237.
442 16. The GAIN Collaborative Research Group. New models of collaboration in genome-
443 wide association studies: the Genetic Association Information Network. Nat Genet.
444 2007; **39**: 1045–1051.

17. Sanders AR, Duan J, Levinson DF, Shi J, He D, Hou C, et al. No Significant Association of 14 Candidate Genes With Schizophrenia in a Large European Ancestry Sample: Implications for Psychiatric Genetics. Am J Psychiatry. 2008; **165**: 497–506.

18. Dick DM, Foroud T, Flury L, Bowman ES, Miller MJ, Rau NL, et al. Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative. Am J Hum Genet. 2003; **73**: 107–114.

19. Kassem L, Lopez V, Hedeker D, Steele J, Zandi P, Bipolar Disorder Consortium NIMH Genetics Initiative, et al. Familiality of Polarity at Illness Onset in Bipolar Affective Disorder. Am J Psychiatry. 2006; **163**: 1754–1759.

20. Green EK, Raybould R, Macgregor S, Hyde S, Young AH, O'Donovan MC, et al. Genetic variation of brain-derived neurotrophic factor (BDNF) in bipolar disorder. Br J Psychiatry. 2006; **188**: 21–25.

21. Green EK, Raybould R, Macgregor S, Gordon-Smith K, Heron J, Hyde S, et al. Operation of the Schizophrenia Susceptibility Gene, Neuregulin 1, Across Traditional Diagnostic Boundaries to Increase Risk for Bipolar Disorder. Arch Gen Psychiatry. 2005; **62**: 642.

22. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006; **35**: 34–41.

23. 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**: 661–678.

24. 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**: 1217–1224.

25. 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**: 1182–1192.

26. Monson ET, Pirooznia M, Parla J, Kramer M, Goes FS, Gaine ME, et al. Assessment of whole-exome sequence data in attempted suicide within a bipolar cohort. Eur Neuropsychopharmacol. 2017; **3**: 1–11.

27. McInnis MG, Dick DM, Willour VL, Avramopoulos D, MacKinnon DF, Simpson SG, et al. Genome-wide scan and conditional analysis in bipolar disorder: evidence for genomic interaction in the National Institute of Mental Health genetics initiative bipolar pedigrees. Biol Psychiatry. 2003; **54**: 1265–1273.

28. 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.

29. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Hum Genet. 2014; **133**: 639–650.

30. Boraska V, Franklin CS, Floyd JAB, Thornton LM, Huckins LM, Southam L, et al. A genome-wide association study of anorexia nervosa. Mol Psychiatry. 2014; **19**: 1085–1094.

31. 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**: 185–190.

32. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010; **466**: 368–372.

33. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of

Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. Am J Hum Genet. 2014; **94**: 677–694.

34. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotech. 2011; **29**: 512–520.

35. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009; **460**: 753–757.

36. Bergen SE, O'Dushlaine CT, Ripke S, Lee PH, Ruderfer DM, Akterin S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. Mol Psychiatry. 2012; **17**: 880–886.

37. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Project: Schizophrenia Trial Design and Protocol Development. Schizophr Bull. 2003; **29**: 15–31.

38. Sullivan PF, Lin D, Tzeng J-Y, van den Oord E, Perkins D, Stroup TS, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008; **13**: 570–584.

39. Ware JJ, Chen X, Vink J, Loukola A, Minica C, Pool R, et al. Genome-Wide Meta-Analysis of Cotinine Levels in Cigarette Smokers Identifies Locus at 4q13.2. Sci Rep. 2016; **6**: 20092.

40. Hamshere ML, Walters JTR, Smith R, Richards AL, Green E, Grozeva D, et al. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. Mol

Psychiatry. 2013; **18**: 708–712.

41. 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**: 103–112.

42. Guo Y, Wei Z, Keating BJ, Hakonarson H, Nervos GCA, Consor WTCC, et al. Machine learning derived risk prediction of anorexia nervosa. BMC Med Genomics. 2016; **9**: 4.

43. 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.

44. 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**: 515–521.

45. 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**: 4875–4889.

46. 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**: 675–684.

47. 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.