Predictive Machine Learning for Personalised Medicine in Major Depressive Disorder

Viktoria-Eleni Gountouna[1a], Mairead L. Bermingham[1a], Ksenia Kuznetsova[a], Daniel Urda Muñoz[b], Felix Agakov[b], Siân E. Robson[a], Joeri J. Meijsen[a,c], Archie Campbell[a], Caroline Hayward[a], Eleanor M. Wigmore[a,d], Toni-Kim Clarke[d], Ana Maria Fernandez[a,d], Donald J. MacIntyre[d], Paul McKeigue[b,e], David J. Porteous[a], Kristin K. Nicodemus[2e,a]

a. Centre for Genomics and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

b. Pharmatics Ltd., Edinburgh, United Kingdom

c. Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom

d. Division of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom

e. Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom

1. These authors contributed equally to this work.

2. To whom correspondence should be addressed:

Kristin K. Nicodemus, Ph.D., M.P.H.

Usher Institute of Population Health Sciences and Informatics

University of Edinburgh

Edinburgh

United Kingdom

Email: kristin.nicodemus@ed.ac.uk

Tel: +44 (0)74 1919 6637

The authors have declared that no conflict of interest exists.

Classification: Biological Sciences/Medical Sciences

Abstract

Depression is a common psychiatric disorder with substantial recurrence risk. Accurate prediction from easily collected data would aid in diagnosis, treatment and prevention. We used machine learning in the Generation Scotland cohort to predict lifetime risk of depression and, among cases, recurrent depression. Rank aggregation was used to combine results across ten different algorithms and identify highly predictive variables. The model containing all but the cardiometabolic predictors had the highest predictive ability on independent data. Rank aggregation produced a reduced set of predictors without decreasing predictive performance (lifetime: 20 out of 154 predictors and Receiver Operating Characteristic area under the curve (AUC)=0·84, recurrent: 10 out of 180 predictors and AUC=0·76). Here we develop a pipeline which leads to a small set of highly predictive variables. This information can be easily collected with a smartphone 'application' to help diagnosis and treatment, while longitudinal tracking may help patients in self-management.

Keywords: machine learning, prediction, psychiatry, depression

Significance

Depression is the most common psychiatric disorder and a leading cause of disability worldwide. Patients are often diagnosed and treated by non-specialist clinicians who have limited time available to assess them. We present a novel methodology which allowed us to identify a small set of highly predictive variables for a diagnosis of depression, or recurrent depression in patients. This information can easily be collected using a tablet or smartphone application in the clinic to aid diagnosis.

Introduction

Major depressive disorder (MDD) is one of the most common mental disorders with a lifetime prevalence of around 15% (1), and it is also frequently recurrent (2-5) or persistent. The World Health Organization predicts that by 2030 13% of the total global disease burden will be accounted for by depression (6). Reducing the burden of MDD is a key public health challenge for the 21st century (7), with accurate diagnosis and prediction of recurrence of critical importance in reducing the disease burden. However, algorithms to predict incident and recurrent MDD that include data above that obtained on the basis of a structured clinical interview have contributed to improved prediction accuracies and may also improve diagnosis (8-10), especially when many patients see non-specialist clinicians such as general practitioners for diagnosis and treatment, where structured clinical interviews are impractical due to time constraints.

Prediction algorithms using standard statistical methodology have been developed previously for incident and/or recurrent MDD (7-12). These algorithms generated nominal to fairly high discriminative accuracy up to a C-statistic or AUC value of 0·79 (8-10). In these previous studies, the number and type of available predictors were limited to a small number of clinical features and applied standard statistical methodologies to develop prediction algorithms. However, in the age of “Big Data” – large-scale biobanking efforts with large sample sizes and hundreds to thousands of potential predictors – these standard statistical methodologies may be limited in their ability to advance personalized medicine in high-dimensional data. However, for clinical utility it is

crucial to identify a concise and easily measured set of predictors with high predictive ability which can be used to determine a patient's risk for lifetime or recurrent MDD. Ideally, these data could be combined with electronic health records to improve prediction and increase clinical utility. Indeed, advanced approaches such as machine learning have been successfully applied to prediction of treatment outcome in MDD (13).

We applied a set of modern "Big Data" approaches – a set of state-of-the-art machine learning methodologies – to the Generation Scotland: Scottish Family Health Study (GS:SFHS) cohort to predict lifetime and recurrent MDD. The predictors in GS:SFHS include cognitive function, personality, health and family history, socio-demographic, biometric, clinical and genomic data (14-15**).** We sought to apply these algorithms to a set of nested models to assess (a) algorithm performance and (b) to define a sparse set of predictors that may be useful in prediction in clinical practice by using a novel approach: the Markov Chain 4 (MC4) algorithm, which was developed for rank aggregation of Internet search engine rankings (16).

Results

Design Framework

We employed an array of machine learning algorithms, including tree-based, regression-based, neural networks and support vector machines (Methods). Data was divided into training and test sets (Fig. 1a). The training data was used to optimise hyperparameters for each algorithm using ten-fold cross validation and build prediction models (Fig. 1b). We assessed the performance of four nested models, described in detail in the Methods. Rank aggregation was used on the variable importance measures from the training data to select the top variables across methods. We performed independent replication on the test data using the models built on the training data with the area under the ROC curve as an outcome of interest.

Demographics

For lifetime MDD, in both the training and test data, individuals with MDD were significantly younger, more likely to be female, to not be living as a couple, to live alone, to report lower income and live in areas that are more socioeconomically deprived than controls (all *p*-values < Bonferroni-corrected threshold of 0·005; SI Appendix, Table S1). There were no significant demographic differences between single and recurrent cases in the training or test sets using the Bonferroni *p*-value threshold.

Predictive Performance

The predictive performance across the four models and machine learning algorithms for lifetime and recurrent MDD are summarised in Table 1 for lifetime MDD and Table 2 for single versus recurrent MDD. In both study designs, model one performed the poorest, model two increased predictive performance, and model three outperformed all other models. Model four, including cardiometabolic predictors, did not significantly improve upon model three; in fact, for some algorithms the inclusion of these additional variables led to significantly decreased predictive ability. For the best-performing algorithm in lifetime MDD, gradient descent boosting (GDB), model three increased the AUC versus model two from approximately 0·71 to 0·84, which is a more than threefold increase in information for discrimination, from 0.4 bits to 1.4 bits, although AUC values were slightly lower than what is commonly used for biomarkers (17). This improvement was also observed for recurrent MDD, although AUC values for Model two were smaller, less than 0·6. The increase in AUC for Model three for the best model, random forest (RF), increased the AUC from Model two of 0·58 to around 0·76.

After adjusting for 45 tests within each model and focusing on Model three as the best performing model, the GDB AUC was significantly higher than both C5.0 and conditional inference forest (CIF; *p*-values 0·00017 and 0·000028, respectively; SI Appendix, Table S2); other algorithms performed as well as GDB. For single versus recurrent MDD, C5.0

and linear support vector machine (SVM-L) showed significantly poorer performance versus the best-performing algorithms (SI Appendix, Table S3).

Ranked Variable Importance and MC4 Analysis

The MC4 aggregate-ranked variable importance measures were used to calculate AUC for both lifetime and recurrent MDD, selecting ten, 20, 30…N of the ranked list to determine the minimum number of variables required to attain maximum AUC values for prediction of MDD in the independent test set. For lifetime MDD, this value was reached after the addition of the top 20 variables. The top ranked variables included (in order): neuroticism, General Health Questionnaire (GHQ) total score, GHQ depression, GHQ somatic symptoms, age, family history of depression, income, live as a couple, sex, mother with depression, whether the participant owned their home, GHQ anxiety, ever smoked, Mood Disorder Questionnaire (MDQ) score, Schizotypal Personality Questionnaire (SPQ) total score, educational qualification, pain intensity, GHQ social dysfunction, whether they ever had chronic pain, and if they live with someone else or live alone (Fig. 2a). For recurrent MDD, the maximum AUC was achieved after including only ten of the top-ranked variables. The top-ranking variables included (in order of rank): age at MDD onset, neuroticism, GHQ total score, digit symbol substitution (processing speed), GHQ somatic symptoms, GHQ depression, age, GHQ social dysfunction, whether they own their home and age when starting smoking (Fig. 2b). We observed considerable overlap between the two sets of predictors, including GHQ total and subscale depression, somatic symptoms and depression scores, neuroticism, age, and whether the participant owns their home. However, there were interesting

differences, where for lifetime MDD family history of depression, MDQ and SPQ scores were critical predictors that were not included in the models predicting recurrent depression; and for recurrent depression processing speed was ranked highly by MC4, which was not included in the set of predictors for lifetime depression.

A comparison of the MC4-ranked variables and variables ranked by individual algorithms (Fig. 2a and 2b) showed that, with the exception of CIF in recurrent depression, none of the methods were fully consistent with the aggregated ranking. In fact, EN only ranked three predictors in the top 20 for lifetime depression and none in the top ten for recurrent depression. However, using the MC4-ranked subset of 20 variables for lifetime MDD and ten variables for single episode versus recurrent MDD led to equal – if not improved – performance across all methods. In lifetime MDD, the MC4-based model AUC values did not differ from those using the full set of 154 variables (Table 2). In single versus recurrent MDD, the reduced set of ten variables produced from MC4 improved predictive performance at the *p*-value < 0·05 level across six of the ten methods, for example, increasing the AUC for SVM-L from 0·695 using 180 variables in Model three to 0·771 using the MC4 set of ten.

Leave-One-Out Analysis of MC4 Sets

To assess the relative contribution of predictors in the MC4 sets, we performed leave-one-variable-out analysis of these subsets, holding the random number seed constant. For lifetime MDD, only the removal of neuroticism showed a significant reduction in AUC values across machines (Δ AUC range = 0·024-0·034; *p*-values range = 0·00018 –

2·80 x 10^{-08}; Supplementary Table 4). For single versus recurrent MDD, the removal of age at onset showed the largest reduction in AUC values across algorithms, with a change in AUC ranging from 0.052 for CIF to 0.088 for C5.0 (significant *p*-value range = 0·00052 to 2·36x10^{-05}). Other variables showed a variable pattern across methods when being removed; none showed a consistent increase or decrease, which was expected as MC4 aggregates rankings across methods. However, the change in AUC values were small for all other variables.

Effect size of Model 3 and MC4 Models

Hedges *g* was used to estimate effect sizes across different algorithms for Model three and the MC4 model in both study designs (Supplementary Table 6). Lifetime effect sizes were large, with all *g* values > 1·0 (Model three range = 1·06 to 1·41 MC4 range = 1·05 to 1·26). For single versus recurrent MDD, the use of the MC4 set increased *g* values; in fact, for several algorithms the *g* values were within the range of “medium” (e.g., < 0·8) when using Model three, but all *g* values were large (ranging from 0·93 to 1·13) using MC4 (Supplementary Table 6).

Discussion

Using state-of-the-art machine learning methods combined with a novel usage of the MC4 algorithm, we have defined a set of predictive variables with AUC values that are larger than those observed in previous studies; for lifetime MDD, the AUC value is within the range of predictive ability for use in clinic. These algorithms were trained on a large biobank with deep phenotyping, in contrast to previous approaches, and may be easily applied to other similar “Big Data” applications.

The most consistent demographic factors associated with lifetime MDD risk in previous studies have been female sex, younger age and low socioeconomic status (18-22). Psychosocial risk factors commonly associated with MDD include negative life events, traumatic experiences, work-related stress, financial strain, poor marital or interpersonal relationships, lack of social support and low self-esteem (10, 23-26). The number of previous episodes, the level of residual symptoms, and childhood maltreatment are consistent risk factors of recurrent MDD (27-28). We highlight that our findings suggest neuroticism is one of the top predictors of lifetime MDD and also recurrent MDD. Neuroticism is easily measured by a general practitioner in clinic and would be the best single measure for prediction of both types of depression.

Using our algorithm, we provide descriptive characteristics of individuals who are at each quintile of our prediction engine, as an illustration of patients a general practitioner may see in clinic (Table 3). For example, versus individuals who would have the lowest predicted risk of lifetime MDD (Table 4), individuals at higher risk levels would be more

likely to score higher on measures of neuroticism, general psychological distress (GHQ), have a family history of depression, be a current smoker and live alone. For individuals with a higher predicted probability of having recurrent depression versus single-episode (Table 5), higher risk for recurrent MDD would be more likely observed in individuals with a lower age at MDD onset, higher neuroticism, higher psychological distress (GHQ), and poorer performance on digit symbol coding.

The main strength of this study is that it has assessed the importance and contribution of well-known and novel predictors of MDD measured using standardised tests in a large population-based cohort; and further introduces the MC4 algorithm for rank aggregation of results. Generation Scotland is a large homogeneous population that to date is one of the few cohorts containing a clinical definition of MDD on a large-scale sample. The two study designs (lifetime MDD–healthy controls and single episode-recurrent MDD) allow for insights into shared and differing risk factors. Limitations of this study are the sample size and the lack of replication in an independent large-scale cohort with a clinical definition of MDD.

This study has shown that a combination of simple self-report measures can form the basis of models that provide excellent prediction of lifetime or recurrent MDD. Given that the measures identified through the MC4 algorithm are relatively easy to obtain, a possible implementation of the findings of this work could be to collect these measures at regular intervals, which would provide a basis for longitudinal assessment of change in individuals or groups of patients. Such evidence on patient outcomes would be useful

in tracking the impact of interventions, especially treatment response to antidepressants or psychological interventions such as cognitive behavioural therapy. If incorporated into clinical practice, the use of algorithms to predict a patient's risk of MDD could have a substantial impact on the timeliness and effectiveness of diagnosis and treatment. The measures can be obtained outside of the clinic and the resulting data-based risk prediction can inform clinicians' decisions, allowing more time during consultations to discuss issues specific to the patient to accommodate a personalised approach. Another exciting prospect for the implementation of the results of this study could therefore be to use the MC4 algorithm as the basis of a decision support tool that could be incorporated into a mobile application (app) for the benefit of patients, support workers and clinicians. Patients could enter their own data into the app and a built-in algorithm would be able to forewarn or reassure about lifestyle changes that might presage or mitigate recurrence. Apps are increasingly being developed and used successfully in health and social care to access treatment guidelines and to support decisions about patient screening, treatment options and drug dosage (29). In the current era of increased availability of data, new systems for linking and utilising healthcare and other records could provide crucial and timely information to support personalised healthcare.

Materials and Methods

Generation Scotland: Scottish Family Health Study

Our sample was drawn from the GS:SFHS, a large population-based study. A full description of the cohort and protocol for recruitment is described in detail elsewhere (14-15). Briefly, the GS:SFHS is a representative survey of the Scottish population, consisting of 23,960 individuals over 18 years of age recruited between the years 2006 and 2011. All components of GS:SFHS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89) and all participants gave written informed consent. Generation Scotland data are available under managed access by submitting a proposal to the Generation Scotland Data Access Committee.

The selection criteria for participants included: Caucasian ethnicity, born in the UK and phenotype data available from attendance at a Generation Scotland research clinic (30). We further restricted the analysis to unrelated individuals, as the machine learning algorithms employed required independent observations. Cases were not matched to controls as doing so would remove the effect of the matching criteria from analysis; thus preventing assessment of potential interaction effects for those criteria.

Assessment of MDD

All participants in GS:SFHS were screened for psychiatric disorders and those who responded positively to screening questions were invited to continue with the MDD and bipolar disorder modules from the Structured Clinical Interview for Diagnostic and

Statistical Manual of Mental Disorders, Fourth Edition (SCID;31). The diagnostic interview was conducted in person by trained clinical research nurses.

Predictors of MDD: Self-Report, Clinical, Cognitive and Demographic

MDD Symptomology

For analysis of single versus recurrent depression, we included 12 symptom predictors from the SCID interview, plus age at onset and a variable indicating which nurse conducted the interview. Missing values for SCID interview questions were controlled for by creating a binary vector indicating which values were not recorded.

General Psychological Distress

Psychological distress was measured using the 28-item General Health Questionnaire (GHQ; 32) at baseline by a trained clinical nurse. Fifty-seven percent of the study participants were also given the Mood Disorder Questionnaire (MDQ; 33) and Schizotypal Personality Questionnaire-Brief (SPQ-B; 34).

Cognitive Performance

Cognitive performance was assessed using measures of processing speed (Wechsler Digit Symbol Substitution Task; 35), verbal declarative memory (Wechsler Logical Memory Test; 36) both immediate and delayed, executive function measured with the

letter-based phonemic verbal fluency test using the letters C, F, and L, each for one minute (37) and vocabulary (Mill Hill Vocabulary Scale; 38).

Demographic and Socioeconomic Status

The Scottish Index of Multiple Deprivation (SIMD; 39) 2009, matched to each participant's postcode, was used to assess socioeconomic status. SIMD is a ranking based on seven domains: income, employment, health, education, geographic access, crime, and housing.

Clinical Measurements, Personal and Family Medical History

Clinical measures included height, weight, BMI, waist, hip, waist:hip ratio, percent body fat, systolic and diastolic blood-pressure (both mean of two measurements), respiratory function (forced expiratory volume in one second, forced vital capacity and forced expiratory flow; each measured three times) and blood sodium, potassium, urea, creatinine, glucose, total cholesterol and HDL-cholesterol levels (14). Participants were asked about whether they or their first-degree relatives had a history of depression, Parkinson disease, Alzheimer disease, diabetes, asthma, cardiovascular disease (including heart disease, hypertension and stroke), osteoarthritis, rheumatoid arthritis, hip fracture or common cancers (including breast, bowel, lung and prostate).

Alcohol and Tobacco Use

Participants were identified as current drinkers, former drinkers (either stopping greater than or less than 12 months) or never drinkers. Current alcohol consumption was characterised as either more or less than usual for each participant. Total consumption was measured in self-reported units of alcohol consumed in the previous week. Participants were classified into two groups: ever-smokers (current smokers and former smokers) and never smokers, with ever-smokers also asked at what age they started smoking and amount of tobacco consumption.

Chronic Pain

A validated self-report chronic pain questionnaire was used to assess pain severity based on its intensity and impact of pain on daily functioning in the previous three months[40].

Analytic Methods

Training and Test Datasets

One thousand and nineteen randomly-selected unrelated participants who had at least one episode of major depressive disorder were randomly selected as cases. Randomly-selected, healthy unrelated individuals (N = 3,994) with no lifetime diagnosis of psychiatric disorders served as controls. For analyses examining single versus recurrent major depressive disorder, we randomly resampled unrelated individuals from

GS:SFHS to increase sample size (as controls were excluded), leading to a total of 1,198.

The final data sets were divided into (a) training data for machine learning hyper-parameter optimisation and model building and (b) test data (Fig. 1a). The test data were not used to develop any prediction models and formed an independent replication set. Many machine learning methods for binary classification require relatively equal numbers of cases and controls to assure optimised prediction (41). We performed a down-sampling procedure to balance the case-control ratio in the training data. The minority class (e.g., cases with MDD and recurrent MDD) were randomly split into a training set (63%) and an independent test set (37%). The majority class (e.g., controls and single episode MDD) was randomly under-sampled to make its frequency equal to the minority class for the training data (see Table 1 for Ns). Hedges g (42) was used to estimate effect sizes because of imbalance in the independent test sets.

Machine Learning Algorithms and Hyper-Parameter Settings

A panel of predictive machine learning algorithms was applied to the data. We applied the C5.0 tree-based algorithm (43-44), two tree-based ensemble algorithms (random forest, RF; 45-46) and conditional inference forest (CIF; 47-48), a gradient descent boosting algorithm (GDB; 49-50), elastic net, a regularised regression technique (EN; 51-52), feed-forward neural networks (NN; 53-54), and support vector machines (SVM; 55-57) with the following kernels: linear (SVM-L), polynomial (SVM-P) and radial basis function (SVM-R). We also used forward stepwise regression with Akaike's Information

Criterion (58). These methods were selected as best examples of a range of machine learning methodologies commonly applied to biologic data.

The R package caret (59) was used to perform ten-fold cross validation on the training data to set optimised values for the hyper-parameters for each algorithm (Fig. 1b; details about hyper-parameter selection in SI Appendix, Methods). All analyses were conducted in R version 3.2.4 (60).

Receiver Operating Characteristic (ROC) area under the curve (AUC) and Variable Importance Measures (VIMs)

The outcome of interest was the area under the ROC curve from the independent test data using models built on the training data. We used the nonparametric DeLong (61) method as implemented in pROC (62) to test for statistical differences between areas under the curve among independent test data. We sought to find the largest ROC AUC whilst minimising the number of variables required such that a smaller set could potentially be used for prediction by general practitioners. We also report an alternative measure to the AUC to aid in interpretation, where to quantify differences in predictive performance we use the expected weight of evidence as measured in bits (63). Machine learning algorithms generate measures, or VIMs, of the relative contribution of predictors in classifier construction, or a measure of the strength of association between predictor and outcome. For C5.0 the VIM used was the percent of observations in the nodes directly under the split for that variable; for RF and CIF it was the permutation-

based VIM; for EN and FSR it was the absolute value of the coefficients, for NN and all SVM we used caret's internal VIM function, which is not dependent on the algorithm and simply measures the AUC increase when including the variable over a null model as these methods do not provide a direct measure of variable importance.

Markov Chain 4 (MC4) Algorithm

We took ranked variable importance measures from the training data using the optimal set of hyper-parameters for each algorithm and applied the Markov Chain 4 (MC4) algorithm (16, 64) to provide an aggregated variable ranking. To begin, the MC4 algorithm considers a single set U that contains the union of the top k elements from all lists. For each pair i and j in the list ($i \neq j$), set the Markov chain transition matrix M element m_{ij} to $1/|U|$ if $\geq$ 50% of the lists rank j above i and $0/|U|$ otherwise. If i and j are never contained in the same list, then $m_{ij} = m_{ji} = 0.5/|U|$, and let $m_{ii} =$. If ε is a small positive constant, it creates an ergodic transition matrix M by $((1- \varepsilon)*m) + \varepsilon/|U|$. The resulting stationary probabilities are then used to create an aggregate ranking of the variables, with higher probabilities denoting a higher rank (i.e., the Markov process spent longer time in those states). The proposed novel use of meta-ranking technology provides robust inferences about the relative contribution to risk for MDD across different machines. We then selected the top ten, 20, …, N ranked variables and fitted them in the best-performing model to assess the number of variables required to reach maximum predictive ability. Leaving one predictor out at a time, we assessed the

relative contribution of each predictor to the final model AUC values in the independent test set.

Models Assessed

We assessed performance of four nested models, where *outcome*$_k$ was either (1) lifetime MDD versus controls or (2) single versus recurrent depression:

Model 1: $Outcome_k = \beta_1 age + \beta_2 age^2 + \beta_3 sex$

Model 2: $Outcome_k = \beta_1 age + \beta_2 age^2 + \beta_3 sex + \sum_{i=4}^{18} \beta_i sociodemographic\ variables$

Building upon model two, the third model included mood, schizotypy and cognitive variables; family and personal medical history (excluding stroke, metabolic and cardiovascular history); as well as smoking and alcohol consumption. In addition, for the analysis of single versus recurrent MDD, we added variables derived from the SCID interview; this led to a total of 154 variables for lifetime MDD and 180 for analysis of single versus recurrent MDD. We hypothesised that cardiometabolic predictors may not improve prediction in all MDD cases but instead particularly in those with a later age at onset (65). Model four included Model three plus cardiovascular and stroke medical and family history; obesity, diabetes, FEV and lab measurements as these are commonly co-morbid with depression (N predictors: lifetime MDD: 211; single versus recurrent MDD: 237).

Hyperparameters were optimised using ten-fold cross validation on the training data. We used the nonparametric DeLong approach to test for statistical differences between areas under the curve among independent test data.

Author Contributions

KKN designed the study, supervised and performed analysis, and wrote the manuscript. VEG, MLB, and KK performed analysis and wrote the manuscript. JJM and SER wrote the manuscript. DJP co-designed the study and edited the manuscript. DUM, FA, PM, AC, CH, EW, TC, AMF, DJM performed analysis or assistance in data coding and edited the manuscript. AM edited the manuscript.

Acknowledgments

VEG and KKN were supported by a University of Edinburgh-Medical Research Council (https://mrc.ukri.org/) Confidence in Concept award and by a Wellcome Trust (https://wellcome.ac.uk/)-University of Edinburgh Institutional Strategic Support Fund award (to KKN). KKN, MLB, DUM, FA, PM, DJP were supported by a University of Edinburgh-Medical Research Council (https://mrc.ukri.org/) Confidence in Concept award (to DJP). KKN and JJM were supported by the Rosetrees Trust (http://www.rosetreestrust.co.uk/; M405) and VEG, SER and KKN by a Chancellor's Fellowship from the University of Edinburgh. DUM was supported by European Union FP7 (https://ec.europa.eu/research/fp7/index_en.cfm) 316861 MLPM. DJM acknowledges the financial support of NHS Research Scotland (http://www.nhsresearchscotland.org.uk/; NRS), through NHS Lothian. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (http://www.cso.scot.nhs.uk/) [CZD/16/6] and the Scottish Funding Council (http://www.sfc.ac.uk/) [HR03006]. Genotyping of the

GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK (https://mrc.ukri.org/) and the Wellcome Trust (https://wellcome.ac.uk/; Wellcome Trust Strategic Award “STratifying Resilience and Depression Longitudinally” (STRADL) Reference 104036/Z/14/Z). We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Kessler RC, et al. (2005) Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* 62(6):593.

2. Hardeveld F, Spijker J, De Graaf R, Nolen WA, Beekman ATF (2013) Recurrence of major depressive disorder and its predictors in the general population: results from The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Psychol Med* 43(01):39–48.

3. Mueller TI, et al. (1999) Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *Am J Psychiatry* 156(7):1000–6.

4. Wang J (2004) A longitudinal population-based study of treated and untreated major depression. *Med Care* 42(6):543–550.

5. Wang JL, Patten SB, Currie S, Sareen J, Schmitz N (2012) Predictors of 1-year outcomes of major depressive disorder among individuals with a lifetime diagnosis: a population-based study. *Psychol Med* 42(02):327–334.

6. Sayers J (2001) The world health report 2001 - Mental health: new understanding, new hope. *Bull World Health Organ* 79:1085–1085.

7. King M, et al. (2008) Development and Validation of an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice Attendees. *Arch Gen Psychiatry* 65(12):1368.

8. van Loo HM, Aggen SH, Gardner CO, Kendler KS (2015) Multiple risk factors predict recurrence of major depressive disorder in women. *J Affect Disord* 180:52–61.

9. Wang JL, et al. (2013) Development and validation of prediction algorithms for major depressive episode in the general population. *J Affect Disord* 151(1):39–45.

10. Wang JL, et al. (2014) DEVELOPMENT AND VALIDATION OF A PREDICTION ALGORITHM FOR USE BY HEALTH PROFESSIONALS IN PREDICTION OF RECURRENCE OF MAJOR DEPRESSION. *Depress Anxiety* 31(5):451–457.

11. King M, et al. (2013) Predicting onset of major depression in general practice attendees in Europe: extending the application of the predictD risk algorithm from 12 to 24 months. *Psychol Med* 43(09):1929–1939.

12. Wang J, et al. (2014) A prediction algorithm for first onset of major depression in the general population: development and validation. *J Epidemiol Community Health* 68(5):418–24.

13. Chekroud AM, et al. (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 3(3):243–250.

14. Smith BH, et al. (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* 42(3):689–700.

15. Smith BH, et al. (2006) Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 7(1):74.

16. Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the Web. *Proceedings of the Tenth International Conference on World Wide Web - WWW '01* (ACM Press, New York, New York, USA), pp 613–622.

17. Xia J, Broadhurst DI, Wilson M, Wishart DS (2013) Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* 9(2):280–299.

18. Skapinakis P, Weich S, Lewis G, Singleton N, Araya R (2006) Socio-economic position and common mental disorders. *Br J Psychiatry* 189(02):109–117.

19. Wang JL, Schmitz N, Dewa CS (2010) Socioeconomic status and the risk of major depression: the Canadian National Population Health Survey. *J Epidemiol Community Health* 64(5):447–52.

20. Weich S, Lewis G (1998) Material standard of living, social class, and the prevalence of the common mental disorders in Great Britain. *J Epidemiol Community Health* 52(1):8–14.

21. Weich S, Lewis G (1998) Poverty, unemployment, and common mental disorders: population based cohort study. *BMJ* 317(7151):115–9.

22. Weich S, Sloggett A, Lewis G (1998) Social roles and gender difference in the prevalence of common mental disorders. *Br J Psychiatry* 173(06):489–493.

23. BLAZER DG, HYBELS CF (2005) Origins of depression in later life. *Psychol Med* 35(09):1241.

24. GOLDBERG D (2006) The aetiology of depression. *Psychol Med* 36(10):1341.

25. Hardeveld F, Spijker J, De Graaf R, Nolen WA, Beekman ATF (2010) Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr Scand* 122(3):184–191.

26. Hope S, Power C, Rodgers B (1999) Does financial hardship account for elevated psychological distress in lone mothers? *Soc Sci Med* 49(12):1637–1649.

27. Wang J, Patten SB, Currie S, Sareen J, Schmitz N (2012) A Population-based Longitudinal Study on Work Environmental Factors and the Risk of Major Depressive Disorder. *Am J Epidemiol* 176(1):52–59.

28. Nanni V, Uher R, Danese A (2012) Childhood Maltreatment Predicts Unfavorable Course of Illness and Treatment Outcome in Depression: A Meta-Analysis. *Am J Psychiatry* 169(2):141–151.

29. NHS Education for Scotland Mobile Knowledge.

30. Raven JC, Court JH (1977) *Manual for Raven's Progressive Matrices and Vocabulary Scales* (HK Lewis).

31. SIMD (2009) *Scottish Index of Multiple Deprivation*.

32. Kerr SM, et al. (2013) Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Med Genet* 14(1):38.

33. First M (2005) *Structured clinical interview for DSM-IV-TR axis I disorders: patient edition* (Biometrics Research Department, Columbia University).

34. Goldberg DP, Hillier VF (1979) A scaled version of the General Health Questionnaire. *Psychol Med* 9(01):139.

35. Hirschfeld RMA (2002) The Mood Disorder Questionnaire: A Simple, Patient-Rated Screening Instrument for Bipolar Disorder. *Prim Care Companion J Clin Psychiatry* 4(1):9–11.

36. Raine A, Benishay D (1995) The SPQ-B: A Brief Screening Instrument for Schizotypal Personality Disorder. *J Pers Disord* 9(4):346–355.

37. Wechsler D (1998) *WAIS-III UK Wechsler Adult Intelligence Scale* (Psychol Corp).

38. Wechsler D (1998) *WAIS-III UK Wechsler Memory Scale-Revised* (Psychol Corp).

39. Lezak M, Howieson D, Bigler E, Tranel D (2012) *Neuropsychological Assessment* (Oxford University Press).

40. Von Korffa M, Ormela J, Keefeb FJ, Dworkinc SF (1992) Grading the severity of chronic pain. *Pain* 50(2):133–149.

41. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans Syst Man, Cybern Part C (Applications Rev* 42(4):463–484.

42. Ridgeway G (2010) gbm: generalized boosted regression models. R package version 1.6-3.1.

43. Hedges L V. (1981) Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *J Educ Stat* 6(2):107–128.

44. Quinlan JR (John R (1993) *C4.5 : programs for machine learning* (Morgan Kaufmann Publishers).

45. Kuhn M, Weston S, Coulter N, Culp M (2015) C5.0 decision trees and rule-based models for pattern recognition. R package version 0.1.0-24.

46. Breiman L (2001) Random Forests. *Mach Learn* 45(1):5–32.

47. Wright MN (2017) ranger: A fast implementation of Random Forests. R package version 0.8.0.

48. Hothorn T, Hornik K, Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *J Comput Graph Stat* 15(3):651–674.

49. Hothorn T, Hornik K, Strobl C, Zeileis A (2017) party: A laboratory for recursive partytioning. R package version 1.2-3.

50. Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat* 29:1189–1232.

51. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33(1):1–22.

52. Friedman J, Hastie T, Simon N, Qian J, Tibshirani R (2017) glmnet: Lasso and elastic-net regularized generalized linear models. R package version 2.0-13.

53. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44(3):837.

54. Robin X, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12(1):77.

55. Ripley BD (1996) *Pattern Recognition and Neural Networks* (Cambridge University Press).

56. Ripley B (2016) nnet: Feed-forward neural networks and multinomial log-linear models. R package version 7.3-12.

57. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297.

58. David Meyer D, et al. (2017) e1071: Misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien. R package version 1.6-8.

59. Karatozoglou A, Smola A, Hornik K (2016) kernlab: Kernel-based machine learning lab. R package version 0.9-25.

60. Ripley B, et al. (2017) Support functions and datasets for Venables and Ripley's MASS. R package 7.3-47.

61. Kuhn M (2017) caret: Classification and regression training. R package 6.0-77.

62. R Core Team (2016) R: A Language and Environment for Statistical Computing. Available at: https://www.r-project.org/.

63. McKeigue P (2017) Sample size requirements for learning to classify with high-dimensional biomarker panels. *Stat Methods Med Res*:096228021773880.

64. Slawski M, Boulesteix A-L (2017) GeneSelector: Stability and aggregation of ranked gene lists. R package version 2.28.0.

65. Seldenrijk A, et al. (2011) Carotid atherosclerosis in depression and anxiety: Associations for age of depression onset. *World J Biol Psychiatry* 12(7):549–558.

Fig. 1

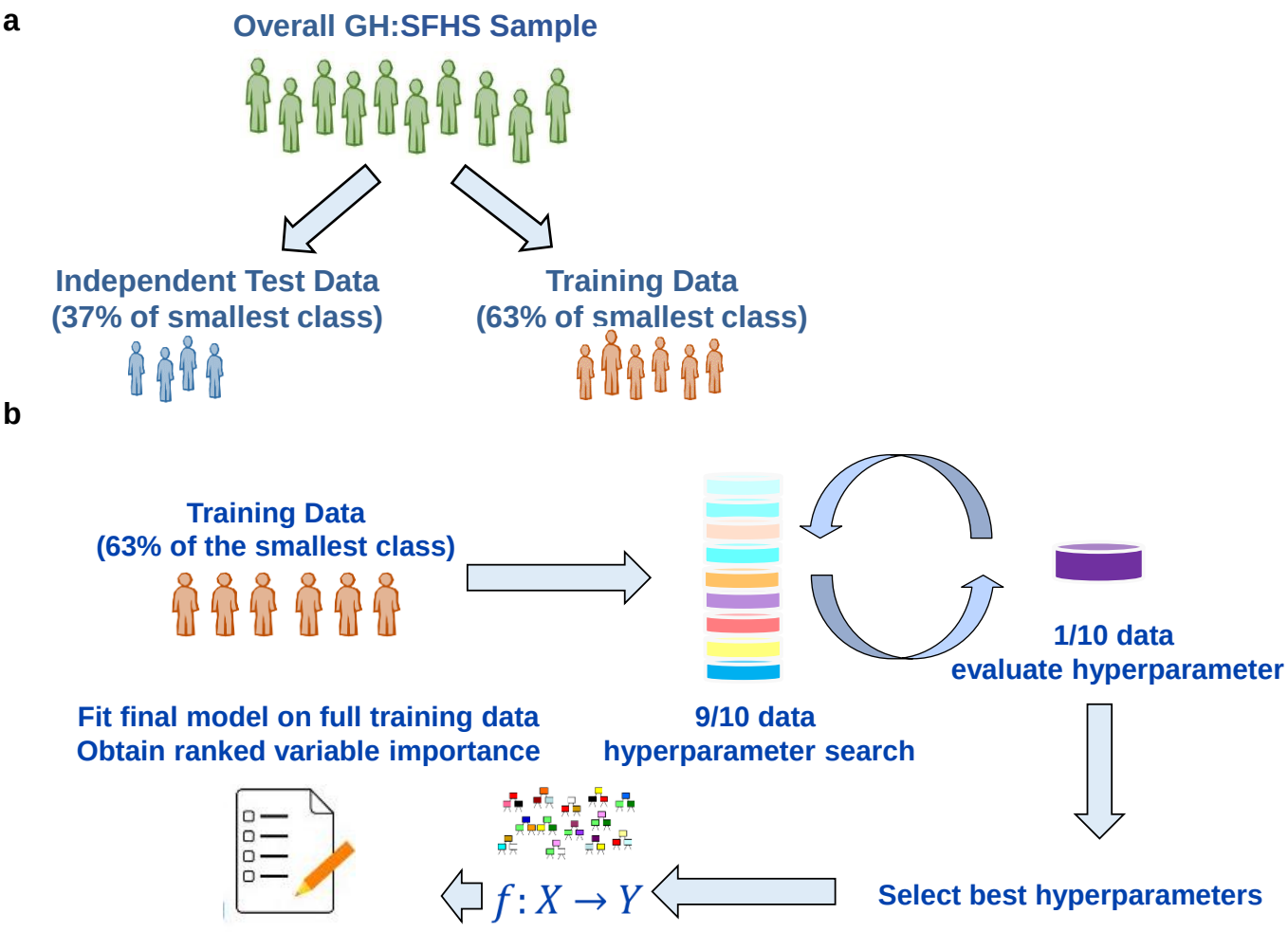
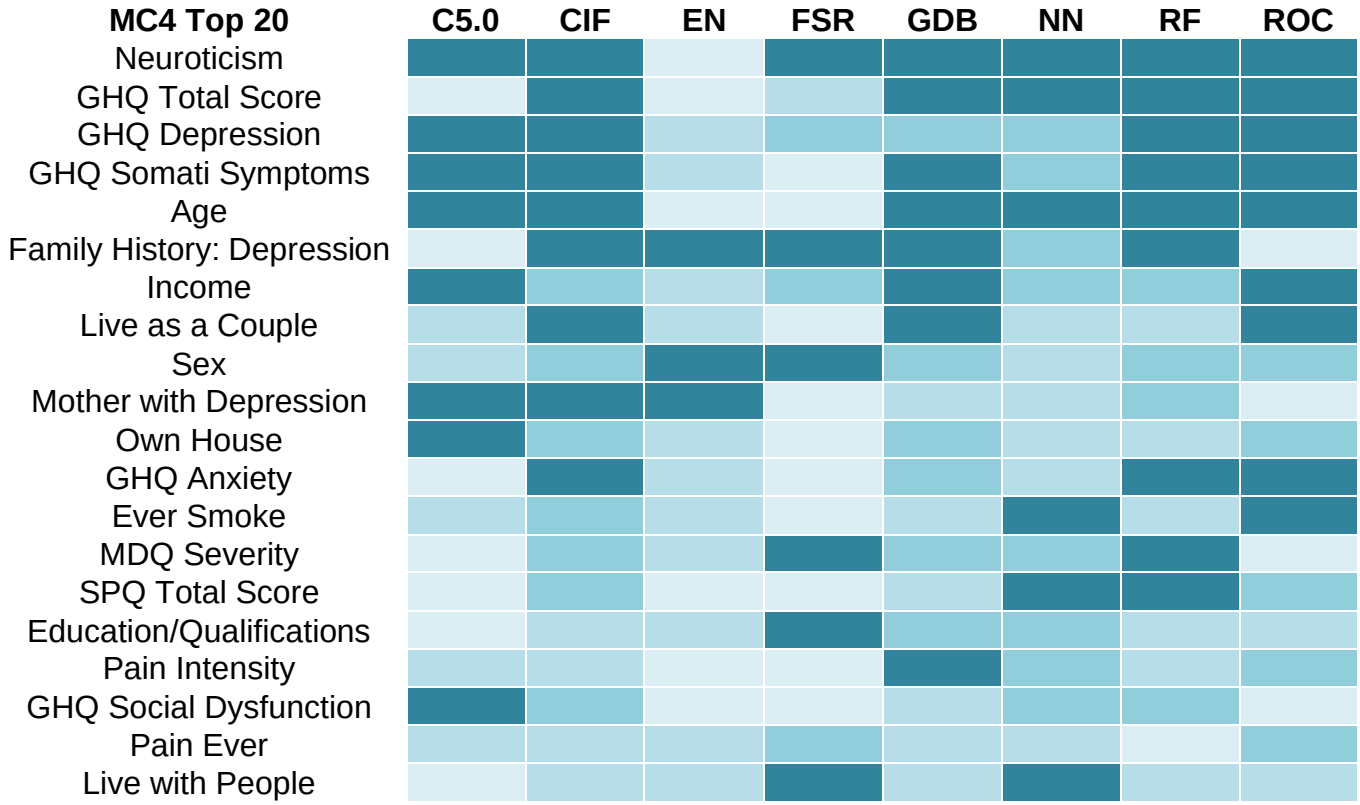


Fig. 1 Schematic of Study Design. (a). Training and test sample creation. (b). Overall schematic of the study design. First, the training data are put through 10-fold cross validation to estimate appropriate hyperparameters. Second, the correct hyperparameters are applied to the full training set to create a model for prediction, and the variable importance measure rankings are recorded. Finally, the models developed on the training data are used for prediction of either lifetime MDD or single versus recurrent MDD on the test data.

Fig. 2

a



b

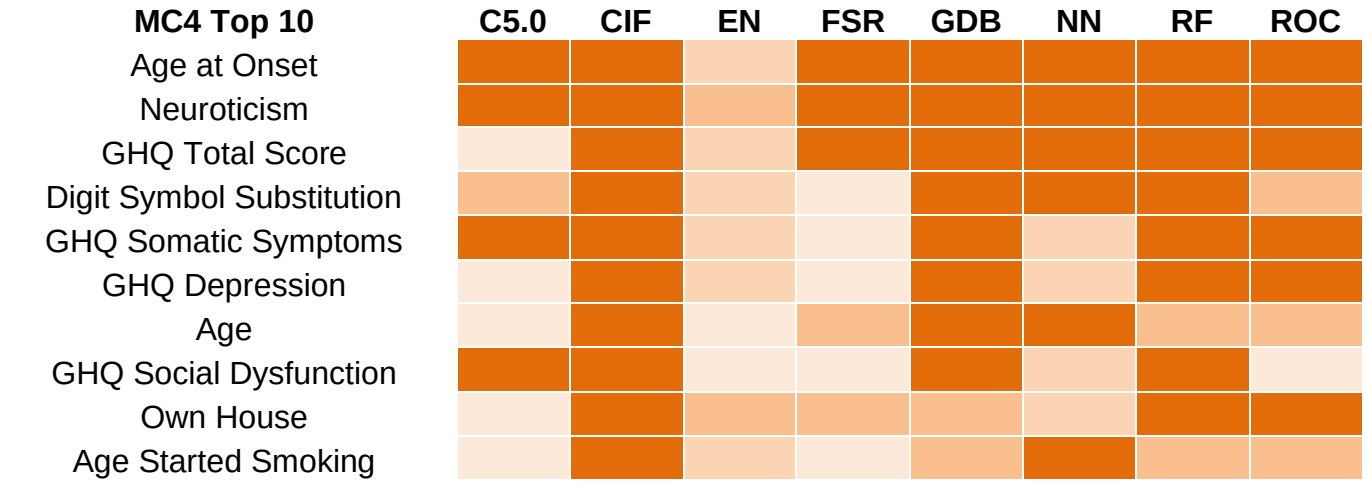


Fig. 2 Concordance of individual algorithm rankings versus MC4 overall ranking. (a) Top panel shows concordance in lifetime MDD. Darkest blue = individual methods ranked in top 10, next darkest blue = top 20, light blue = top half, lightest blue = bottom half. The number of predictors in C5.0 (52) and FSR (41) was less than the top half of predictors, so any predictor with a zero-valued VIM was set to dark blue. GHQ = General Health Questionnaire, MDQ = Mood Disorder Questionnaire, SPQ = Schizotypal Personality Questionnaire (b) Bottom panel shows concordance in single versus recurrent episode MDD. Darkest orange = individual methods ranked in top 10, next darkest orange = top 20, light orange = top half, lightest orange = bottom half. The number of predictors in C5.0 (40), EN (37) and FSR (30) was less than the top half of predictors, so any predictor with a zero-valued VIM was set to dark orange. GHQ = General Health Questionnaire.

Table 1. AUC Results for Predicting MDD on Independent Test Data: Lifetime MDD

Algorithm	M1 AUC	M1 95% CI	M2 AUC	M2 95% CI	M3 AUC	M3 95% CI	M4 AUC	M4 95% CI
C5.0	0·652	(0·625, 0·679)	0·694*	(0·666, 0·723)	0·813**	(0·789, 0·837)	0·805	(0·780, 0·830)
Conditional inference forest	0·670	(0·642, 0·697)	0·709*	(0·681, 0·737)	0·825****	(0·802, 0·848)	0·829*	(0·806, 0·852)
Elastic net	0·670	(0·643, 0·697)	0·705*	(0·677, 0·734)	0·835****	(0·812, 0·858)	0·834	(0·811, 0·857)
Forward stepwise regression	0·669	(0·642, 0·697)	0·711*	(0·683, 0·739)	0·829****	(0·805, 0·853)	0·820	(0·795, 0·844)
Gradient descent boosting	0·663	(0·635, 0·69)	0·712**	(0·683, 0·741)	**0·839*** **	**(0·816, 0·861)**	0·840	(0·819, 0·862)
Neural networks	0·666	(0·640, 0·694)	0·709*	(0·681, 0·736)	0·824****	(0·801, 0·847)	0·825	(0·801, 0·849)
Random forest	0·660	(0·632, 0·688)	0·710*	(0·681, 0·739)	0·824****	(0·801, 0·848)	0·828	(0·804, 0·851)
SVM, linear	0·666	(0·639, 0·694)	0·704*	(0·675, 0·734)	0·833****	(0·809, 0·856)	0·830	(0·807, 0·854)
SVM, polynomial	0·670	(0·639, 0·694)	0·704*	(0·675, 0·733)	0·830****	(0·809, 0·856)	0·831	(0·807, 0·854)
SVM, radial basis function	0·652	(0·623, 0·680)	0·694*	(0·665, 0·722)	0·829****	(0·806, 0·853)	0·830	(0·806, 0·853)

Table 2. AUC Results for Predicting MDD on Independent Test Data: Single versus Recurrent MDD

Algorithm	M1 AUC	M1 95% CI	M2 AUC	M2 95% CI	M3 AUC	M3 95% CI	M4 AUC	M4 95% CI
C5.0	0·555	(0·518, 0·592)	0·536	(0·487, 0·585)	0·689**	(0·644, 0·734)	0·650*	(0·584, 0·681)
Conditional inference forest	0·593	(0·543, 0·644)	0·603	(0·554, 0·653)	0·728**	(0·686, 0·770)	0·742*	(0·701, 0·783)
Elastic net	0·584	(0·532, 0·636)	0·601	(0·552, 0·650)	0·742**	(0·700, 0·784)	0·739	(0·696, 0·782)
Forward stepwise regression	0·589	(0·537, 0·640)	0·570	(0·520, 0·619)	0·699*	(0·655, 0·742)	0·651*	(0·604, 0·698)
Gradient descent boosting	0·592	(0·541, 0·643)	0·600	(0·549, 0·648)	0·756***	(0·715, 0·796)	0·735*	(0·692, 0·778)
Neural networks	0·587	(0·538, 0·636)	0·557	(0·507, 0·606)	0·749***	(0·706, 0·792)	0·713*	(0·668, 0·757)
Random forest	0·544	(0·495, 0·594)	0·576	(0·526, 0·626)	**0·759*** **	**(0·718, 0·800)**	0·743	(0·700, 0·787)
SVM, linear	0·587	(0·535, 0·639)	0·592	(0·543, 0·641)	0·695*	(0·650, 0·740)	0·667*	(0·619, 0·715)
SVM, polynomial	0·584	(0·531, 0·636)	0·595	(0·546, 0·644)	0·709*	(0·665, 0·754)	0·697	(0·652, 0·742)
SVM, radial basis function	0·557	(0·507, 0·606)	0·571	(0·522, 0·621)	0·705**	(0·660, 0·750)	0·692	(0·646, 0·738)

M = Model, AUC = Receiving Operator Characteristic Area Under the Curve, CI = Confidence Interval. The best performing model is denoted in bold type. * $p<0.05$, ** $p<1x10^{-05}$,*** $p<1x10^{-10}$, **** $p<1x10^{-15}$ versus previous model.

Table 3. AUC Results for Best Performing Model versus MC4 Model

	Lifetime MDD: M3 = 154 Variables, MC4 = 20 Variables				Recurrent MDD: M3 = 180 Variables, MC4 = 10 Variables			
Algorithm	**M3 AUC**	**M3 95% CI**	**MC4 AUC**	**MC4 95% CI**	**M3 AUC**	**M3 95% CI**	**MC4 AUC**	**MC4 95% CI**
C5.0	0·813	(0·789, 0·837)	0·808	(0·784, 0·833)	0·689	(0·644, 0·734)	0·735*	(0·694, 0·776)
Conditional inference forest	0·825	(0·802, 0·848)	0·825	(0·802, 0·848)	0·728	(0·686, 0·770)	0·743*	(0·702, 0·784)
Elastic net	0·835	(0·812, 0·858)	0·831	(0·809, 0·854)	0·742	(0·700, 0·784)	0·771*	(0·733, 0·809)
Forward stepwise regression	0·829	(0·805, 0·853)	0·834	(0·812, 0·857)	0·699	(0·655, 0·742)	0·771*	(0·733, 0·809)
Gradient descent boosting	**0·839**	**(0·816, 0·861)**	**0·839**	**(0·817, 0·862)**	0·756	(0·715, 0·796)	**0·775***	**(0·736, 0·813)**
Neural networks	0·824	(0·801, 0·847)	0·835	(0·812, 0·858)	0·749	(0·706, 0·792)	0·752	(0·712, 0·792)
Random forest	0·824	(0·801, 0·848)	0·830	(0·808, 0·853)	**0·759**	**(0·718, 0·800)**	0·762	(0·721, 0·803)
SVM, linear	0·833	(0·809, 0·856)	0·832	(0·809, 0·855)	0·695	(0·650, 0·740)	0·771*	(0·732, 0·809)
SVM, polynomial	0·830	(0·809, 0·856)	0·832	(0·809, 0·854)	0·709	(0·665, 0·754)	0·740	(0·698, 0·782)
SVM, radial basis function	0·829	(0·806, 0·853)	0·827	(0·805, 0·850)	0·705	(0·660, 0·750)	0·743	(0·700, 0·786)

M = Model, AUC = Receiving Operator Characteristic Area Under the Curve, CI = Confidence Interval. The best performing models are denoted in bold type. * $p<0.05$, ** $p<1x10^{-05}$,*** $p<1x10^{-10}$, **** $p<1x10^{-15}$ for comparison of model 3 to MC4 model.

Table 4. Examples of Individuals at Quintiles of Predicted MDD using MC4 Predictors: Lifetime MDD

Prediction Quintile	Neu	GHQ Total	GHQ-Dep	GHQ-Som	Age	Fx Dep	Income Level	Live as Couple	Sex	Mother Depress	Own House	GHQ-Anx	Ever Smoke	MDQ-C	SPQ-Total	Ed/Qual	Pain Inten	GHQ-Soc	Pain Ever	Number in House
0%	0	7	0	0	65	No	Lowest	Yes	M	No	Yes	0	No	None	1	School Leavers Certificate	40	7	No	2
25%	0	5	0	1	61	No	Highest	No	M	No	Yes	0	Current	None	5	University	38	4	Yes	1
50%	4	14	0	5	64	No	Second-Highest	Yes	F	No	Yes	2	No	None	7	University	21	7	Yes	2
75%	4	11	0	2	49	No	Highest	No	M	No	Yes	6	No	Serious	3	University	56	3	Yes	3
100%	8	22	4	5	25	Yes	Highest	No	F	No	No	10	Current	Serious	12	University	33	3	Yes	1

Table 5. Examples of Individuals at Quintiles of Predicted MDD using MC4 Predictors :Single versus Recurrent MDD

Prediction Quintile	Age Onset	Neu	GHQ-Total	Digit Symbol	GHQ-Som	GHQ-Dep	Age	GHQ-Soc	Own House	Age Smoking
0%	51	0	13	63	1	1	53	8	Yes	14
25%	38	9	27	81	9	0	51	5	Yes	12
50%	30	4	18	67	8	0	41	7	Yes	22
75%	36	10	35	77	4	8	55	15	Yes	17
100%	14	12	56	39	10	12	54	16	No	7

Supporting Information

Supplementary Methods

Machine learning optimisation

The optimisation was set to a grid search for C5.0 (number of trials 1-20), CIF and RF (number of trees set to 1000, number of variables selected at each split (*mtry*) set to a grid of 10-15 values up to the total number of variables), GDB (interaction depth grid set to 1-3, number of iterations grid set to between 1 000 and 5 000, shrinkage held constant at 0.001), EN (grid for alpha and lambda set between 0 and 1). For NN and SVM models, we used a random search as the number of hyper-parameters was large and a random search has been shown to have equal or improved performance versus a grid search for these methods (1). Data were centred and scaled before analysis using SVMs as different scales of measurement can lead to numerical difficulties during the calculation of the inner products of the variables.

Imputation

For variables with less than 5% missing data, multivariate imputation by chained equations (MICE; 2-3) was used to replace missing values.

Supporting Information References

1. Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. *J Mach Learn Res* 13(Feb):281–305.

2. van Buuren S, Groothuis-Oudshoorn K, van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 045(i03).

3. van Buuren S, et al. (2017) mice: Multivariate imputation by chained equations. R package version 2.46.0.

Table S1. Demographic and Socioeconomic Characteristics in Cases and Controls and Single vs. Recurrent MDD

Outcome	Lifetime MDD				Recurrent vs Single MDD			
Data	Training		Test		Training		Test	
Sample Size	642	628	377	3862	274	273	160	588
Group	Case	Control	Case	Control	Recurrent	Single	Recurrent	Single
Scottish nationality (N, %)	580 (90.3)	550 (87.6)	343 (91.0)	3366 (87.2)	257 (93.8)	251 (91.9)	544 (92.5)	146 (91.3)
Age (mean, SD)	49.6 (11.8)	56.7 (11.9)	50.4 (11.7)	56.9 (11.7)	47.2 (12.3)	49.9 (12.4)	48.7 (12.0)	51.8 (12.4)
Female (N, %)	469 (73.1)	359 (57.2)	257 (68.2)	2162 (56.0)	202 (73.7)	195 (71.4)	417 (70.9)	106 (66.3)
Living as a couple (N, %)	381 (59.3)	499 (79.5)	228 (60.5)	3026 (78.4)	166 (60.6)	180 (65.9)	326 (55.4)	106 (66.3)
Living alone (N, %)	154 (24.0)	82 (13.1)	84 (22.3)	545 (14.1)	59 (21.5)	58 (21.2)	149 (25.3)	26 (16.3)
Annual personal income								
£0-£10 000	113 (17.6)	50 (8.0)	59 (15.6)	273 (7.1)	44 (16.1)	34 (12.5)	114 (19.4)	20 (12.5)
£10 001-£30 000	243 (37.9)	181 (28.8)	127 (33.7)	1242 (32.2)	93 (33.9)	96 (35.2)	224 (38.1)	57 (35.6)
£30 001-£50 000	126 (19.6)	169 (26.9)	95 (25.2)	987 (25.6)	65 (23.7)	76 (27.8)	115 (19.6)	32 (20.0)
£50 001 +	126 (19.6)	175 (27.9)	82 (21.8)	1002 (25.6)	59 (21.5)	54 (19.8)	114 (19.4)	38 (23.8)
Refused	34 (5.3)	53 (8.4)	14 (3.7)	358 (9.3)	13 (4.7)	13 (4.8)	21 (3.6)	13 (8.1)
Worked in last 12 months (N, %)	393 (61.2)	420 (66.9)	242 (64.2)	2497 (64.7)	180 (65.7)	182 (66.7)	362 (61.6)	97 (60.6)
Educational Qualification								
Less than secondary	68 (10.6)	66 (10.5)	46 (12.2)	426 (11.0)	31 (11.3)	25 (9.2)	64 (10.9)	17 (10.6)
Secondary	176 (27.4)	212 (33.8)	101 (26.8)	1171 (30.3)	79 (28.8)	79 (28.9)	163 (27.7)	46 (28.8)
Tertiary	398 (62.0)	350 (55.7)	230 (61.0)	2265 (58.6)	164 (59.9)	169 (61.9)	361 (61.4)	97 (60.6)
SIMD								
1 - Most deprived	111 (17.3)	60 (9.6)	70 (18.6)	401 (10.4)	65 (23.7)	37 (13.6)	107 (18.2)	18 (11.3)
2	112 (17.4)	73 (11.6)	63 (16.7)	440 (11.4)	40 (14.6)	36 (13.2)	111 (18.9)	23 (14.4)
3	116 (18.1)	107 (17.0)	53 (14.1)	15.5 (599)	43 (15.7)	48 (17.6)	98 (16.7)	32 (20.0)
4	144 (22.4)	148 (23.6)	87 (23.1)	1024 (26.5)	63 (23.0)	64 (23.4)	125 (21.3)	40 (25.0)
5 - Least deprived	159 (24.8)	240 (38.2)	104 (27.6)	1398 (36.2)	63 (23.0	88 (32.2)	147 (25.0)	47 (29.4)

Yellow: $p<0.005$, light green: $p<0.01$, dark green: $p<0.05$. Statistical significance was tested between cases/controls for lifetime MDD or between single versus recurrent MDD using χ^2 tests for categorical and *t*-tests for continuous outcomes.

Table S2. ROC Curve Comparisons between Methods within Models for Lifetime MDD

(a) Model 1 (above diagonal) and Model 2 (below diagonal)

	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
C5.0		0.019	0.17	0.036	0.17	0.068	0.33	0.13	0.010	0.98
CIF	0.029		0.84	0.96	0.050	0.97	0.018	0.32	0.92	0.0066
EN	0.0086	0.59		0.80	0.062	0.25	0.028	0.26	0.59	0.0055
FSR	0.014	0.78	0.041		0.80	0.43	0.037	0.12	0.87	0.012
GDB	0.0086	0.51	0.20	0.81		0.27	0.20	0.41	0.068	0.11
NN	0.065	0.94	0.54	0.61	0.55		0.13	0.91	0.29	0.021
RF	0.020	0.91	0.36	0.81	0.58	0.86		0.20	0.031	0.21
SVM-L	0.17	0.46	0.46	0.0032	0.14	0.38	0.28		0.31	0.030
SVM-P	0.14	0.38	0.49	0.0032	0.11	0.38	0.15	0.92		0.0064
SVM-R	0.94	0.024	0.067	0.0079	0.0082	0.024	0.0073	0.12	0.043	

(b) Model 3 (above diagonal) and Model 4 (below diagonal)

	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
C5.0		0.12	0.0026	0.062	**0.00017**	0.21	0.17	0.014	0.028	0.039
CIF	**0.00081**		0.039	0.62	**0.000028**	0.86	0.89	0.24	0.31	0.31
EN	**0.0011**	0.41		0.26	0.37	0.080	0.035	0.41	0.072	0.15
FSR	0.18	0.29	0.019		0.18	0.52	0.56	0.43	0.85	0.95
GDB	**8.65E-08**	**0.00010**	0.22	0.012		0.012	0.0068	0.27	0.094	0.085
NN	0.026	0.59	0.13	0.48	0.027		0.95	0.22	0.37	0.45
RF	0.0089	0.69	0.22	0.37	0.014	0.75		0.20	0.16	0.32
SVM-L	0.0081	0.87	0.35	0.030	0.13	0.44	0.70		0.54	0.45
SVM-P	0.0037	0.71	0.16	0.10	0.13	0.44	0.70	0.85		0.81
SVM-R	0.0082	0.94	0.15	0.13	0.054	0.49	0.66	0.88	0.52	

p-values < 0.05 after Bonferroni correction are in bold.

Table S3. ROC Curve Comparisons between Methods within Models for Single versus Recurrent MDD

(a) Model 1 (above diagonal) and Model 2 (below diagonal)

	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
C5.0		0.041	0.13	0.086	0.049	0.095	0.65	0.10	0.17	0.93
CIF	0.0062		0.28	0.57	0.28	0.68	0.017	0.46	0.27	0.015
EN	0.020	0.85		0.47	0.14	0.88	0.11	0.43	0.96	0.11
FSR	0.19	0.016	0.071		0.56	0.92	0.070	0.71	0.31	0.062
GDB	0.0095	0.52	0.89	0.023		0.75	0.033	0.36	0.26	0.027
NN	0.46	0.023	0.049	0.47	0.037		0.079	0.99	0.87	0.14
RF	0.074	0.084	0.26	0.75	0.11	0.42		0.079	0.10	0.57
SVM-L	0.031	0.24	0.42	0.025	0.57	0.047	0.39		0.32	0.083
SVM-P	0.014	0.33	0.66	0.046	0.74	0.032	0.27	0.73		0.13
SVM-R	0.098	0.036	0.15	0.90	0.086	0.45	0.79	0.17	0.036	

(b) Model 3 (above diagonal) and Model 4 (below diagonal)

	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
C5.0		0.050	0.0052	0.67	**0.00017**	0.0055	**0.00019**	0.78	0.36	0.47
CIF	**1.85E-07**		0.24	0.18	0.0071	0.17	0.0053	0.10	0.36	0.25
EN	**4.82E-07**	0.81		0.0086	0.17	0.64	0.18	**0.00092**	0.037	0.018
FSR	0.54	**0.00045**	**0.00012**		0.0022	0.020	0.0036	0.80	0.57	0.74
GDB	**1.62E-06**	0.54	0.72	**0.00058**		0.58	0.68	**0.00025**	0.0068	0.0023
NN	**0.00054**	0.059	0.10	0.021	0.14		0.47	0.0049	0.038	0.024
RF	**2.26E-07**	0.88	0.74	**0.00030**	0.22	0.047		**0.00029**	0.0022	**0.00038**
SVM-L	0.20	**0.00031**	**4.91E-06**	0.45	**0.00024**	0.032	0.032		0.14	0.34
SVM-P	0.014	0.020	0.023	0.077	0.054	0.47	0.014	0.054		0.55
SVM-R	0.024	0.0094	0.0075	0.096	0.022	0.34	0.0043	0.054	0.52	

p-values < 0.05 after Bonferroni correction are in bold.

Table S4. AUC Differences between MC4 Full Model and Leave-One-Out Analysis of Top 20 Predictors in Lifetime Depression

MC4 Top 20	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
Neuroticism	0.034 *(0.00018)*	0.032 *($3.56e^{-05}$)*	0.030 *($2.80e^{-08}$)*	0.031 *($1.08e^{-07}$)*	0.027 *($1.63e^{-05}$)*	0.030 *($2.63e^{-07}$)*	0.025 *($6.02e^{-06}$)*	0.031 *($4.27e^{-08}$)*	0.028 *($2.04e^{-07}$)*	0.024 *($3.07e^{-05}$)*
GHQ Total Score	0.004	-0.001	0.004	-0.004	-0.002	-0.010	-0.007	0.0009	-0.002	-0.004
GHQ Depression	-0.004	0.0009	0.0008	-0.0002	0.003	-0.010	-0.004	0.005	-0.006	0.0005
GHQ Somatic Symptoms	0.0009	-0.003	0.004	-0.002	0	-0.0099	0	0.002	-0.002	-0.0003
Age	0.008	0.003	0.009	0.0002	0.008	-0.006	0.002	0.006	0.003	0.001
Family History: Depression	-0.003	0.003	0.007	0.002	0.002	-0.006	-0.003	0.004	0.0008	-0.0003
Income	0.001	-0.002	0.003	-0.004	-0.001	-0.0098	-0.007	-0.0004	-0.004	-0.004
Live as a Couple	0.008	0.0004	0.006	-0.002	0.001	-0.0084	-0.005	0.003	-0.002	-0.003
Sex	0.001	-0.002	0.004	-0.003	0	-0.010	-0.007	0.002	-0.001	-0.003
Mother with Depression	-0.007	-0.002	0.002	-0.002	0	-0.011	-0.006	0.001	-0.002	-0.003
Own House	-0.003	-0.002	0.004	-0.001	0	-0.0096	-0.007	0.002	-0.002	-0.003
GHQ Anxiety	0.001	-0.002	0.004	-0.002	-0.001	-0.010	-0.002	0.0001	-0.003	-0.004
Ever Smoke	0.0006	-0.002	0.006	-0.002	0.001	-0.0085	-0.005	0.001	-0.0001	-0.002
MDQ Severity Score	-0.002	0.0005	0.004	-0.002	0	-0.0087	-0.004	0.002	-0.002	-0.002
SPQ Total Score	-0.0004	-0.003	0.005	-0.002	-0.001	-0.011	-0.007	-0.002	-0.0004	-0.006
Education/Qualifications	0.005	0.002	0.006	-0.002	0.002	-0.0088	-0.003	0.002	-0.0005	-0.003
Pain Intensity	-0.0007	-0.002	0.005	-0.002	0	-0.0093	-0.007	0.001	-0.002	-0.003
GHQ Social Dysfunction	-0.003	-0.002	0.004	-0.002	-0.001	-0.010	-0.0008	0.001	-0.002	-0.003
Pain Ever	-0.002	-0.003	0.002	-0.005	-0.001	-0.012	-0.006	-0.001	-0.007	-0.006
Live with People	0.001	-0.001	0.003	-0.004	0	-0.010	-0.007	0.0001	-0.003	-0.005

p-values included in parentheses for variables associated with a significant decrease in AUC values after removal, after Bonferroni correction.

Table S5. AUC Differences between MC4 Full Model and Leave-One-Out Analysis of Top 20 Predictors in Single versus Recurrent Depression

MC4 Top 10	C5.0	CIF	EN	FSR	GDB	NN	RF	SVM-L	SVM-P	SVM-R
Age at Onset	0.088 (*0.00056)*	0.052	0.081 (*4.29E-05*)	0.081 (*4.30E-05*)	0.074 (*5.32E-05*)	0.081 (*0.00042*)	0.068 (*0.00014*)	0.079 (*2.36E-05*)	0.073 (*5.96E-05*)	0.067 (*0.00032*)
Neuroticism	0.023	0.004	0.012	0.012	0.014	0.012	0.014	0.015	0.008	0.018
GHQ Total Score	0	0	0	0	0.003	-0.003	0	0.001	0.007	0.003
Digit Symbol Substitution	-0.020	-0.002	0	0	-0.002	-0.008	-0.004	0.001	-0.009	-0.006
GHQ Somatic Symptoms	0.013	-0.003	-0.001	-0.001	-0.001	0	0.001	0.002	-0.001	0.005
GHQ Depression	-0.006	0.001	-0.001	-0.001	0.003	-0.005	0.001	0.001	-0.004	-0.001
Age	0.031	0.001	0.006	0.015	0.012	0	0.005	0.005	0.002	0.007
GHQ Social Dysfunction	-0.008	0.001	-0.001	-0.001	0.006	-0.009	0.006	0.001	0	0.007
Own House	-0.007	0	0	0	0	-0.006	-0.002	-0.001	-0.032 (*0.00047*)	-0.009
Age Started Smoking	-0.001	0.004	0.003	0.003	0.004	-0.009	0.004	0.005	-0.005	0.006

p-values included in parentheses for variables associated with a significant decrease in AUC values after removal, after Bonferroni correction.

Table S6. Effect Size Given by Hedges *g* for Model 3 and MC4 Model in Lifetime and Single versus Recurrent MDD

	Lifetime MDD: M3 = 154 Variables, MC4 = 20 Variables				Recurrent MDD: M3 = 180 Variables, MC4 = 10 Variables			
Algorithm	**M3 *g***	**M3 *g* 95% CI**	**MC4 *g***	**MC4 *g* 95% CI**	**M3 *g***	**M3 *g* 95% CI**	**MC4 *g***	**MC4 *g* 95% CI**
C5.0	1.26	(1.15, 1.37)	1.23	(1.12, 1.34)	0.70	(0.52, 0.88)	0.93	(0.75, 1.12)
Conditional inference forest	1.20	(1.09, 1.31)	1.14	(1.03, 1.25)	0.87	(0.69, 1.05)	0.99	(0.80, 1.17)
Elastic net	1.41	(1.30, 1.52)	1.26	(1.15, 1.37)	0.94	(0.76, 1.12)	1.13	(0.95, 1.32)
Forward stepwise regression	1.12	(1.01, 1.23)	1.26	(1.15, 1.37	0.62	(0.44, 0.80)	1.13	(0.95, 1.32)
Gradient descent boosting	1.25	(1.14, 1.35)	1.23	(1.12, 1.34)	1.04	(0.85, 1.22)	1.16	(0.98, 1.34)
Neural networks	1.23	(1.12, 1.34)	1.21	(1.10, 1.32)	0.97	(0.78, 1.15)	1.02	(0.85, 1.21)
Random forest	1.10	(0.99, 1.21)	1.05	(0.94, 1.16)	1.02	(0.84, 1.20)	1.04	(0.86, 1.22)
SVM, linear	1.14	(1.03, 1.22)	1.16	(1.06, 1.28)	0.71	(0.53, 0.89)	1.12	(0.93, 1.30)
SVM, polynomial	1.13	(1.02, 1.24)	1.15	(1.05, 1.26)	0.80	(0.62, 0.98)	0.96	(0.77, 1.14)
SVM, radial basis function	1.06	(0.95, 1.17)	1.12	(1.01, 1.23)	0.76	(0.59, 0.94)	0.94	(0.76, 1.12)