Data mining algorithm predicts a range of adverse outcomes in major depression

van Loo, Hanna M.; Bigdeli, Tim B.; Milaneschi, Yuri; Aggen, Steven H.; Kendler, Kenneth S.

Published in:
Journal of Affective Disorders

DOI:
10.1016/j.jad.2020.07.098

Journal of Affective Disorders 276 (2020) 945–953

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Data mining algorithm predicts a range of adverse outcomes in major depression

Hanna M. van Loo[a,*], Tim B. Bigdeli[b,c], Yuri Milaneschi[d], Steven H. Aggen[b], Kenneth S. Kendler[b,e]

[a] Department of Psychiatry, University of Groningen, University Medical Center Groningen, Hanzeplein 1 (PO Box 30.001), 9700 RB Groningen, the Netherlands
[b] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, United States
[c] Department of Psychiatry and Behavioral Sciences, State University of New York Downstate Medical Center, Brooklyn, NY, United States
[d] Department of Psychiatry, Amsterdam Public Health and Neuroscience Amsterdam research institutes, Amsterdam UMC and GGZ inGeest Amsterdam, Amsterdam, the Netherlands
[e] Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, United States

ARTICLE INFO

Keywords:
Major depression
Data mining, prediction
Course of illness
Recurrence
Replication

ABSTRACT

Background: Course of illness in major depression (MD) is highly varied, which might lead to both under- and overtreatment if clinicians adhere to a 'one-size-fits-all' approach. Novel opportunities in data mining could lead to prediction models that can assist clinicians in treatment decisions tailored to the individual patient. This study assesses the performance of a previously developed data mining algorithm to predict future episodes of MD based on clinical information in new data.

Methods: We applied a prediction model utilizing baseline clinical characteristics in subjects who reported lifetime MD to two independent test samples (total n = 4226). We assessed the model's performance to predict future episodes of MD, anxiety disorders, and disability during follow-up (1–9 years after baseline). In addition, we compared its prediction performance with well-known risk factors for a severe course of illness.

Results: Our model consistently predicted future episodes of MD in both test samples (AUC 0.68–0.73, modest prediction). Equally accurately, it predicted episodes of generalized anxiety disorder, panic disorder and disability (AUC 0.65–0.78). Our model predicted these outcomes more accurately than risk factors for a severe course of illness such as family history of MD and lifetime traumas.

Limitations: Prediction accuracy might be different for specific subgroups, such as hospitalized patients or patients with a different cultural background.

Conclusions: Our prediction model consistently predicted a range of adverse outcomes in MD across two independent test samples derived from studies in different subpopulations, countries, using different measurement procedures. This replication study holds promise for application in clinical practice.

1. Introduction

The course of major depression (MD) can be highly varied (Eaton et al., 2008), which may lead to either over- or undertreatment in clinical practice if a generic treatment regimen is adopted. Data mining techniques offer opportunities to develop prediction algorithms for clinically relevant outcomes such as course of illness (Hastie et al., 2009). Data mining uses pattern recognition techniques to extract important patterns and trends from data, for example with the aim to predict outcomes. If sufficiently accurate, the resulting prediction models could assist clinicians in identifying patients with a distinct course of illness, and thus support more specific treatment allocation (Darcy et al., 2016), for instance on decisions whether to continue or discontinue treatment after recovery of MD. In different medical disciplines, these opportunities are now being explored (Jiang et al., 2017), in order to move from a 'one-size-fits-all' approach to treatment assignments that are more tailored to the individual patient.

Also in psychiatry, scientists have started to leverage data mining for patient care. Several previous studies developed algorithms to predict course-related outcomes in MD, such as recurrence of MD (van Loo et al., 2015; Wang et al., 2014), treatment resistance or remission (Chekroud et al., 2016; de Vries et al., 2018; Perlis, 2013), or episode persistence, chronicity, hospitalization and disability (van Loo et al., 2014a; Wardenaar et al., 2014). However, most studies were limited by the use of relatively few predictors (van Loo et al., 2014a; Wardenaar et al., 2014), cross-sectional data (van Loo et al.,

* Corresponding author.
E-mail address: h.van.loo@umcg.nl (H.M. van Loo).

https://doi.org/10.1016/j.jad.2020.07.098
Received 18 March 2020; Received in revised form 15 June 2020; Accepted 5 July 2020
Available online 21 July 2020

2014a; Wardenaar et al., 2014), questionnaire instead of interview-based clinical data (van Loo et al., 2014a; Wang et al., 2014; Wardenaar et al., 2014), a short follow-up period (Chekroud et al., 2016; de Vries et al., 2018; Perlis, 2013), or data from women only (van Loo et al., 2015).

In a recently published study, we developed a prediction model for recurrence of MD in an attempt to address these previous limitations (van Loo et al., 2018). We used prospective data, mostly derived from structured clinical interviews, from a sample of 653 participants who reported an episode of MD in the last year. We used a broad range of clinical characteristics assessed at baseline to optimally predict future episodes of MD. The resulting prediction algorithm model showed promising prediction performance in the training data.

Before implementation in clinical practice, prediction models need to be evaluated in new data, preferably in multiple samples representing the target population, i.e. patients who recovered from MD (Hastie et al., 2009; Perlis, 2013). It is crucial to determine whether estimates of prediction performance are reliable and replicable, as estimates derived from initial training data might be overly optimistic due to overfitting (i.e., the model capitalizes on idiosyncratic features of the training data) (Hastie et al., 2009). The primary aim of this study is to validate our previously developed multivariate prediction model in new data. Our primary research question is: how accurately does our previously developed model predict future episodes of MD in two independent test samples? In addition, we test how well the model predicts a broader set of course-related outcomes (future episodes of generalized anxiety disorder (GAD), panic disorder, disability). This replication study is a necessary step towards implementation in clinical practice.

2. Method

2.1. Previously developed prediction model for recurrence of MD

2.1.1. Training sample

In data mining, multiple independent samples are commonly used to train and test a prediction model. The 'training sample' is used to train or discover a prediction model describing the relation between the predictors and the outcome. Then, this model is tested using new data, the 'test sample', to obtain reliable estimates of prediction performance (see also Supplemental methods).

We previously developed a prediction model for recurrence of MD based on a large number of clinical characteristics at baseline (van Loo et al., 2018), using training data from a longitudinal study of male-male and male-female twin pairs from the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD). This training sample included 653 male and female twins who reported an episode of MD (DSM-III-R) in the year prior to baseline interview, and who were also participating in the follow-up interview which was carried out at least one year later. All participants reported a period of >60 days of (partial) remission or recovery (Frank, 1991), in order to focus on MD recurrence instead of chronicity. To minimize recall bias, we used data from participants who reported a MD episode in the *last year* rather than lifetime (Supplemental Methods). This selection was done to increase the quality of reports about the specific symptoms during the episode of MD, the duration, and other severity indices, which we expected to be higher for participants who recently experienced an episode of MD, than participants who had an episode more than one year ago.

2.1.2. Model discovery

In this study, we analysed a total of 70 potential risk factors using Cox models with elastic net regularization (R-package *glmnet*) to predict the outcome recurrence of MD using time-to-event data. Regularized regression methods include a penalty for model complexity. This penalty results in the selection of predictors via the shrinkage of weaker predictor beta-coefficients towards zero. Regularized methods are useful for studies examining large numbers of predictors as it reduces overfitting and yields sparser models (Hastie et al., 2009).

The 70 risk factors covered characteristics of the (1) recent depressive episode (e.g., specific depressive and anxiety symptoms, level of impairment), (2) current state (age, symptom level), (3) psychiatric history, (4) family history, (5) personality, (6) early and (7) recent adversity, and (8) current social and economic environment.

The elastic net penalty controlled the selection and effect sizes of predictors to increase prediction performance and model interpretation (Zou and Hastie, 2005). The final model was selected based on minimal prediction error as assessed in 10-fold cross-validation (Friedman et al., 2010; Simon et al., 2011). This model retained 24 out of the 70 initial predictors and was highly multifactorial including diverse risk factors such as comorbid anxiety symptoms and disorders, maternal MD, and childhood traumas (Supplemental Table 1). Prediction performance in the training sample was good (AUC~0.75), but the model was not evaluated in independent test data, because the sample was relatively small to create test data.

In this previous study, we also studied sex differences in prediction models for recurrence of MD. Since no prominent sex differences were identified, we selected the model built on training data including both sexes. For detailed information on the discovery phase of the prediction model, including its predictors, and their effect sizes, see Supplemental Table 1 and van Loo et al. (2018).

2.2. Test prediction model

2.2.1. Test samples

We used two independent test samples from VATSPSUD and the Netherlands Study of Depression and Anxiety (NESDA) to assess the prediction performance and generalizability of the previously derived prediction model (van Loo et al., 2018). These studies were selected because of their longitudinal designs and high-quality data: data were primarily based on structured interviews administered by trained interviewers, assessed a large set of the risk factors included in the prediction model, and had relatively few missing observations or drop-outs during follow-up. The study designs, samples, and data collection are described in detail in earlier publications (Kendler and Prescott, 2006; Penninx et al., 2008).

The first test sample combined data from the female-female twins (FF, n = 757) and the male-male/male-female twins (MM-MF, n = 1544) from the VATSPSUD study, but who were *not* included in the original training sample used to develop the prediction model. Thus, we created an independent test sample including 2301 Caucasian twins who reported a lifetime episode of MD at baseline assessment (American Psychiatric Association, 1987), and who were re-interviewed at follow-up at least 1 year after baseline interview. Previous studies showed that the VATSPSUD sample is broadly characteristic of the Caucasian general population in the USA in terms of demographic features and rates of psychopathology (Kendler and Prescott, 2006).

The second test sample was drawn from the Netherlands Study of Depression and Anxiety (NESDA). NESDA is a longitudinal cohort study including 2981 subjects from the Dutch general population, primary care, and specialized mental health care, aged 18–65 at baseline assessment (2004–2007). From this sample, we included 1925 subjects who reported a lifetime episode of MD (American Psychiatric Association, 2000) at baseline, and who were re-interviewed approximately 2, 4, 6 or 9 years after baseline (waves 3, 4, 5, and 6).

All participants provided written informed consent, and the studies were approved by Institutional Review Boards of VCU and VU University Medical centre (Kendler et al., 2008; Penninx et al., 2008).

2.2.2. Assessment and imputation of predictors

Most predictors retained in the prediction model (Table 1) were

Table 1
Characteristics of the training and test samples.

	VATSPSUD training[1]	VATSPSUD test[2]	NESDA test
Sample size (n)	653	2301	1925
Demographics			
Study Origin	USA	USA	Netherlands
Predominant ancestry (%)	Caucasian	Caucasian	N-European
Female sex (%)	34.6[3]	53.2[3]	68.6
Age at interview (μ, SD)	35.2 (8.8)	34.9 (8.6)	42.0 (12.4)
Years of education (μ, SD)	13.1 (2.5)	13.4 (2.5)	11.9 (3.2)
Type of sample			
General population (%)	100	100	18.4
Primary care (%)	0	0	36.2
Specialized mental health care (%)	0	0	45.4
History of MD			
Age at onset (μ, SD)	24.8 (11.1)	22.1 (8.8)	28.4 (12.7)
Number of lifetime episodes (μ, SD)	7.1 (25.8)	4.8 (13.6)	4.6 (9.4)
≥ 2 lifetime episodes (%)	45.6	53.4	48.0
≥ 3 lifetime episodes (%)	33.2	34.4	36.7
≥ 4 lifetime episodes (%)	26.7	23.3	28.2
Lifetime comorbidity			
GAD (%)	18.4	14.0	34.6
Alcohol dependence (%)	43.3	26.7	65.4
Risk score (μ, SD)[4]	1.13 (0.46)	0.75 (0.35)	1.16 (0.42)

GAD, generalized anxiety disorder; μ, mean; MD, major depression; SD, standard deviation; USA, United States of America.
[1]The VATSPSUD training sample included data from 653 male-male/male-female twins (MM-MF) who reported an episode of MD (DSM-III-R) in the year before baseline interview, and whose MD status in the year prior to follow-up interview was known. We selected participants with a last year episode of MD only, in order to reduce recall problems. For further details, we refer to Van Loo et al. 2018.[16]
[2]The VATSPSUD test sample included data from the female-female twins (FF, n = 757) and the male-male/male-female twins (MM-MF, n = 1544) who reported a *lifetime* episode of MD at baseline interview (FF1/MM-MF1). The 653 subjects included in the training sample with a *last year* MD-episode at MF1 were excluded from this test sample.
[3]The percentage of women in the VATSPSUD training sample was relatively low because this sample was drawn from a study of male-male and male-female twins.[17]
[4]Recurrence risk score is based on prediction model for recurrence of MD (see Methods and Supplemental Table 1).

assessed at baseline in both test samples. All 24 predictors were present in VATSPSUD; 21 of 24 predictors were available in NESDA (Supplemental Table 1). Some predictors were assessed with different instruments in NESDA. In these cases, items that were most equivalent to the predictors used in VATSPSUD were selected, and if needed, transformed or categorized to increase comparability. All predictors were assessed at baseline, except for four predictors which were assessed during follow-up in part of the participants. These predictors concerned childhood sexual abuse in NESDA and VATSPSUD-FF; and maternal MD, lifetime GAD, low marital satisfaction in VATSPSUD-MM-MF. As these predictors concerned retrospective reports, or were assessed at baseline in the majority of the sample, we decided not to exclude these predictors in order not to bias the prediction performance downward.

Missingness on most predictors was limited: on average 2.8% of the values in VATSPSUD, and 9.8% in NESDA were missing (Supplemental Table 1). Values for missing predictors were multiply imputed in 10 datasets using Multivariate Imputation by Chained Equations (R-package *mice*, 20 iterations) (van Buuren and Groothuis-Oudshoorn, 2011). All predictors needed in the prediction model were included in these imputations; variables concerning lifetime diagnoses of panic disorder and social phobia were used in NESDA as extra predictors in the imputation to improve imputation results (Carpenter and Kenward, 2013).

2.2.3. Risk score for recurrence of MD

First, we applied the prediction model for recurrence of MD to 10 imputed datasets to create 10 risk scores for each subject. The risk scores were constructed as the sum of the subject's risk factor values multiplied by the corresponding beta weight of that risk factor (as estimated in the VATSPSUD-training sample, Supplemental Table 1), i.e. the linear predictor or prognostic index (Royston and Altman, 2013). We created a single risk score for each subject by averaging their risk scores from each of the 10 imputed datasets.

Because multiple imputation of missing values will often not be feasible in clinical practice, we performed a sensitivity analysis with a risk score where missing observations were replaced by sample means. The sample means of VATSPSUD and NESDA are provided in Supplemental Table 1. In this case, we created one single risk score for each subject by summing all the subject's predictor values –or sample mean in case the value was missing– multiplied by the predictors' corresponding beta weights. Note that this a conservative approach to missingness which might bias downward predictive power.

2.2.4. Assessment of prediction performance

We selected several outcomes during follow-up to test the predictive performance of the risk score. The primary outcome was any episode of MD during follow-up, since the prediction model was trained to predict this outcome (van Loo et al., 2018). Given that patients recovered from MD are not only at risk of future episodes of MD, but also of anxiety disorders and disability (Lamers et al., 2011; Moffitt et al., 2007) –the presence of which could also inform treatment decisions on monitoring and treatment (e.g., continuation of antidepressant medication after recovery of MD) – we also tested the predictive value of the risk score with secondary outcomes. These concerned GAD and panic disorder (Kessler et al., 2005) and severe disability as assessed by the World Health Organization Disability Assessment Schedule (WHODAS-II). All outcomes were dichotomous (Supplemental Table 2, note that time-to-event data were not available).

All statistical analyses were performed in R (R Core Team, 2017; Wickham, 2009). Logistic regression models were used to estimate the association between the risk score at baseline and the outcomes during follow-up (R-packages *stats, rcompanion*) (Mangiafico, 2017; R Core Team, 2017). Model discrimination was assessed using areas under the receiver operating characteristic curve (AUC, R-package *Epi*) (Carstensen et al., 2017; Royston and Altman, 2013).The AUC is a measure of model discrimination or "separation"- do patients predicted to be at higher risk exhibit higher event rates than those predicted to be at lower risk? (Royston and Altman, 2013)

We derived two values from the AUC to facilitate interpretation of the effect size. The success rate difference (SRD = 2AUC-1) is equal to Somers' D or Kendall's tau and thus interpretable as a correlation coefficient. The number needed to take (NNT = 1/SRD) represents the number that one would need to test to have one more 'success' (i.e. adverse outcome) in the higher risk group than in the lower risk group (Kraemer, 2014).

To assess the absolute risk for different levels of the risk score, both test samples were split in quartiles and the proportion of observed adverse outcomes for each quartile was determined.

2.2.5. Comparison of risk score with other risk factors for severe course of illness

To further validate the prediction model, we assessed whether our risk score outperformed other risk factors of a severe course of MD: measures of genetic risk, environmental risk, and neuroticism.

First, we used three measures reflecting genetic risk for MD, i.e. family history of MD, age at onset, and polygenic risk score for MD. These measures have been shown to be associated with a more severe course of illness (Eaton et al., 2008; Hardeveld et al., 2013a; Kendler et al., 2005; Mistry et al., 2018; Peterson et al., 2016). Family history of MD was assessed as the ratio of the number of first-degree

relatives affected with MD divided by the number of first-degree relatives.

We calculated a polygenic risk score for MD for 1662 NESDA participants whose genome wide association study (GWAS) data were available. We used GWAS summary statistics for MD publicly released by the Psychiatric Genomics Consortium, obtained for a subset of 59,851 cases and 113,154 controls after the exclusion of data from 23andMe (Wray et al., 2018). Furthermore, since NESDA data were part of the meta-analysis, we re-ran the meta-analysis after removal of overlapping data (~3 K samples). LDpred was used to compute polygenic risk scores (Supplemental Methods 1) (Vilhjálmsson et al., 2015).

Second, we used two measures of environmental risk –early or lifetime traumas and childhood sexual abuse– which are also associated with a more severe course of illness (Gopinath et al., 2007; Hardeveld et al., 2013b; Paterniti et al., 2017). Third, we assessed the association between the risk score and the personality trait neuroticism. Neuroticism strongly reflects liability for MD (Jeronimus et al., 2016), and is associated with a more severe course of MD (Xia et al., 2011). For details about how these risk factors were assessed, we refer to Supplemental Table 2.

We calculated Pearson's correlation coefficient between our risk score and these other risk factors, and we determined the AUC's of the other risk factors for all adverse outcomes.

3. Results

3.1. Sample characteristics

On average, participants in the test sample from NESDA had a higher risk score for recurrence of MD at baseline than participants in the VATSPSUD test sample (Table 1). NESDA participants also reported more often lifetime episodes of GAD and alcohol dependence at baseline. This may reflect differences in disease severity due to differences in sample ascertainment. Whereas VATSPSUD is based on birth records of twins in Virginia (Kendler and Prescott, 2006), NESDA sampled from the general population, primary care, and specialized mental health care (Penninx et al., 2008). In addition, NESDA included a large proportion of subjects with *current* depressive or anxiety disorders, whereas in VATSPSUD, 653 cases with *last year/current* MD were excluded from the test sample since they were included in the training sample drawn from VATSPSUD. The larger proportion of participants with current/recent episodes of MD (instead of lifetime) in NESDA might also explain the similarity between the risk score in the test sample from NESDA and the training sample from VATSPSUD, which included exclusively subjects with a last year MD episode (Table 1).

As expected, there were high rates of co-occurrence between the different outcomes (Table 2). Correlations between MD at follow-up and GAD, panic disorder, and severe disability at follow-up ranged between 0.34 and 0.76.

3.2. Prospective prediction performance

In both test samples, the risk score significantly predicted future episodes of MD, and also the other adverse course-related outcomes, viz. episodes of GAD, panic disorder, and disability (Table 3). Despite the fact that the risk score was specifically trained to predict MD recurrence, its associations with future episodes of MD, GAD, panic disorder, and disability were equally strong. A standard deviation (SD) increase in risk score corresponded with a double risk of these adverse outcomes (mean OR = 2.1). In addition, the AUC's for future episodes of MD (range 0.68–0.73) were comparable with AUC's for the other outcomes (range 0.65–0.78) as indicated by the overlapping 95% confidence intervals.

Prediction performance for future episodes of MD in both test samples was comparable to the performance in the training sample (AUC's 0.68–0.73 versus AUC 0.75; confidence intervals were overlapping), indicating that there was little overfitting of the prediction model in the training data. This showed that the predictive performance of this model was not specific to our first study, but that the model also predicted adverse outcomes of MD across samples from different subpopulations, two different countries, in which different measurement procedures were used.

Using sensitivity analyses, we assessed to what extent prediction performance decreased when multiple imputation was not used to construct the risk score, but missing values were replaced by sample means, because in clinical practice multiple imputation will often not be feasible. Prediction performance was very comparable for this alternative risk score, i.e. AUC's were at most 0.01 attenuated (Supplemental Table 3).

Dividing participants in quartiles based on their risk score, subjects in the lower risk groups consistently reported fewer adverse outcomes than individuals in the higher risk groups (Fig. 1). ROC-curves showed that the optimal cutpoint of the risk score (i.e., resulting in the maximum sum of sensitivity and specificity) for the risk score was ~0.8 in VATSPSUD and ~1.0 in NESDA, resulting in a mean sensitivity of 74% (range 65–82) and mean specificity of 60% (range 46–76) across the different outcomes (Supplemental Fig. 1). The mean negative predictive value was 63% (range 16–83) and the mean positive predictive value was 15% (range 4–51). This means that at the optimal cutpoint, the score performs better in detecting the true negatives than the true

Table 2
Co-occurrence of the outcomes.

(a). VATSPSUD

	MD 0–1 year	GAD 0–1 year	Panic disorder 0–1 year
MD 0–1 year		16.9	5.0
GAD 0–1 year	0.76		6.1
Panic disorder 0–1 year	0.46	0.50	

(b). NESDA

	MD 0–9 year	GAD 0–9 year	Panic disorder 1–2 year	WHODAS ≥40 2 year	WHODAS ≥40 6 year
MD 0–9 year		8.0	3.2	6.6	7.2
GAD 0–9 year	0.60		3.8	3.9	4.3
Panic disorder 1–2 year	0.34	0.43		2.2	2.1
WHODAS ≥40 2 year	0.52	0.47	0.26		10.4
WHODAS ≥40 6 year	0.52	0.48	0.24	0.70	

GAD, generalized anxiety disorder; MD, major depression; WHODAS, World Health Organization Disability Assessment Schedule.
Rates of co-occurrence or comorbidity in VATSPSUD (n = 2301) and NESDA (n = 1925) as indexed by tetrachoric correlations (lower diagonal) and odds ratios (upper diagonal).

Table 3
Risk score predicting psychopathology and disability during follow-up.

	Sample	OR[1]	95% CI	AUC	95% CI	SRD[2]	NNT[2]	N[3]	mean[4]
Any MD									
MD 0–1 year[5]	VATSPSUD	2.1	1.8–2.3	0.73	0.69–0.76	0.45	2.2	1930	0.16
MD 0–2 year	NESDA	1.9	1.7–2.2	0.68	0.66–0.71	0.37	2.7	1638	0.46
MD 0–9 year	NESDA	2.3	2.0–2.7	0.72	0.69–0.75	0.44	2.3	1522	0.74
Any anxiety disorder									
GAD 0–1 year[5,6]	VATSPSUD	2.5	2.2–2.9	0.78	0.75–0.81	0.56	1.8	1929	0.13
GAD 0–2 year	NESDA	2.0	1.8–2.4	0.70	0.67–0.74	0.40	2.5	1638	0.13
GAD 0–9 year	NESDA	2.3	2.0–2.6	0.73	0.70–0.75	0.45	2.2	1264	0.34
Panic disorder 0–1 year[5]	VATPSUD	2.0	1.6–2.5	0.72	0.63–0.80	0.43	2.3	701	0.07
Panic disorder 1–2 year[5]	NESDA	1.6	1.4–1.9	0.65	0.61–0.68	0.29	3.4	1638	0.16
Disability									
WHODAS ≥40 2 year	NESDA	2.3	2.0–2.7	0.72	0.69–0.76	0.45	2.2	1091	0.24
WHODAS ≥40 6 year	NESDA	2.3	1.9–2.7	0.73	0.69–0.77	0.46	2.2	862	0.19

AUC, area under the receiving operating characteristic curve; CI, confidence interval; FU, follow-up; GAD, generalized anxiety disorder; MD, major depression; N, number; NNT, number needed to take; OR, odds ratio; SRD, success rate difference; WHODAS, World Health Organization Disability Assessment Schedule.
This table represents the estimated associations between the recurrence risk score at baseline (as described in Supplemental Table 1) and several adverse outcomes during follow-up. The strength of association was tested using logistic regression analyses in which the standardized recurrence risk score ($M = 0$, SD = 1) was the independent variable, and the measures of MD, anxiety, and disability were the dependent variables. All outcomes were binary; psychiatric disorders were coded as "1" if the subject reported at least one episode in the time interval. Disability was coded as "1" if the participant's level of disability was high (WHODAS > 40, corresponding roughly with the top 25% in NESDA). Years indicate the approximate number of years after baseline assessment (e.g., 0–2 years concerns the first two years after baseline, etc.). For ROC-curves see Supplemental Figure 1.
[1]All odds ratios are highly significant with *P*-values $< 3 \times 10^{-9}$ (Bonferroni corrected alpha 0.05/21 = 0.002).
[2] Success rate difference (SRD) equals 2AUC-1 (equal to Somers' D or Kendall's tau and thus interpretable as a correlation coefficient). Number needed to take (NNT) equals 1/SRD, and represents the number one needed to sample from the subgroup with the higher risk to have one more 'success' (i.e. adverse outcome) than the lower risk group.
[3]Number (N) of subjects with available data on the dependent variable.
[4]Proportion of subjects reporting this outcome.
[5]Outcome assessed in the 12 months prior to interview wave(s).
[6]In VATSPSUD, we tested the association of the recurrence risk score with episodes of GAD with a duration of ≥1 month instead of ≥6 months in the year prior to interview, because only 4% of cases reported GAD with a duration ≥6 months, which might limit reliability of estimated associations.

positives in a population. This can be attributed to the relatively low prevalence of some of the outcomes. If outcomes are rare (e.g., panic disorder occurred in only 7% of VATSPSUD participants), diagnostic tests will more often result in false positives, and less often in false negatives. From a clinical standpoint, a negative test result could in this case be more valuable than a positive test result (e.g., a negative test result could support a decision to reduce antidepressant use).

3.3. Comparing the model with other risk factors for severe course of illness

We compared the prediction performance of our risk score with several measures of genetic risk, environmental risk, and neuroticism, which are well-known risk factors for a severe course of MD. We did this to assess to what extent the more complex risk score outperformed these simpler risk factors. Three of these risk factors were included as a predictor in our risk score (family history, traumas, childhood sexual abuse), the other three risk factors (age at onset, polygenic risk score for MD, neuroticism) were not.

All these risk factors for a severe course of illness were significantly correlated with our risk score in the expected direction (Supplemental Table 4, all *P*-values < 0.002). Subjects with a higher risk score tended to have more first-degree relatives with MD, a higher polygenic risk for MD, an earlier age at onset, higher neuroticism scores, and reported a higher number of traumas and a history of childhood sexual abuse. Neuroticism was particularly highly correlated with the risk score ($r \sim 0.5$).

The risk score predicted the outcomes more accurately (AUC's 0.65–0.78) than logistic regression models based on one of the other risk factors. Prediction performance of the following risk factors -family history of MD, age at onset, polygenic risk score for MD, and childhood sexual abuse- were all in the same range (AUC's ~0.5–0.6), and lifetime traumas performed slightly better (AUC's 0.55–0.66) (Table 4). However, the model including neuroticism only predicted the outcomes almost as accurately as our risk score (AUC's 0.64–0.76). The confidence intervals of the AUC's were overlapping for most outcomes, except for episodes of MD and GAD in VATSPSUD (Supplemental Table 5). For these two outcomes, the risk score had a significantly higher AUC. Of note, neuroticism was not included in our risk score. While it was included in the model discovery phase, it was not retained in the elastic net penalized model (van Loo et al., 2018), which could be due to multicollinearity between neuroticism and the other predictors.

Because of the relatively strong prediction performance of neuroticism, we performed *post hoc* analyses to investigate whether neuroticism could further enhance prediction performance of the risk score. We performed an unpenalized Cox regression analysis including both our risk score and neuroticism as independent variables to predict MD recurrence in the training data (VATSPSUD, $n = 653$) (van Loo et al., 2018). In this model, our risk score significantly predicted MD recurrence (HR 2.1, CI 1.8–2.4) but neuroticism's effect attenuated to not significant (HR 0.9, CI 0.8–1.1). The addition of neuroticism to the risk score did not improve prediction performance– AUC's based on this model were similar to or lower than these based on the risk score alone.

4. Discussion

4.1. Principal findings

We tested a data mining algorithm for predicting future episodes of MD in subjects with lifetime MD using baseline clinical characteristics. The model consistently predicted future episodes of MD in two independent test samples, despite differences in sample composition, study design and assessment of predictors. In addition, the model predicted future episodes of GAD, panic disorder, and disability comparably. Furthermore, the algorithm outperformed several known risk factors for a more severe course of illness, viz. measures of genetic risk, and environmental risk. Only neuroticism predicted the adverse

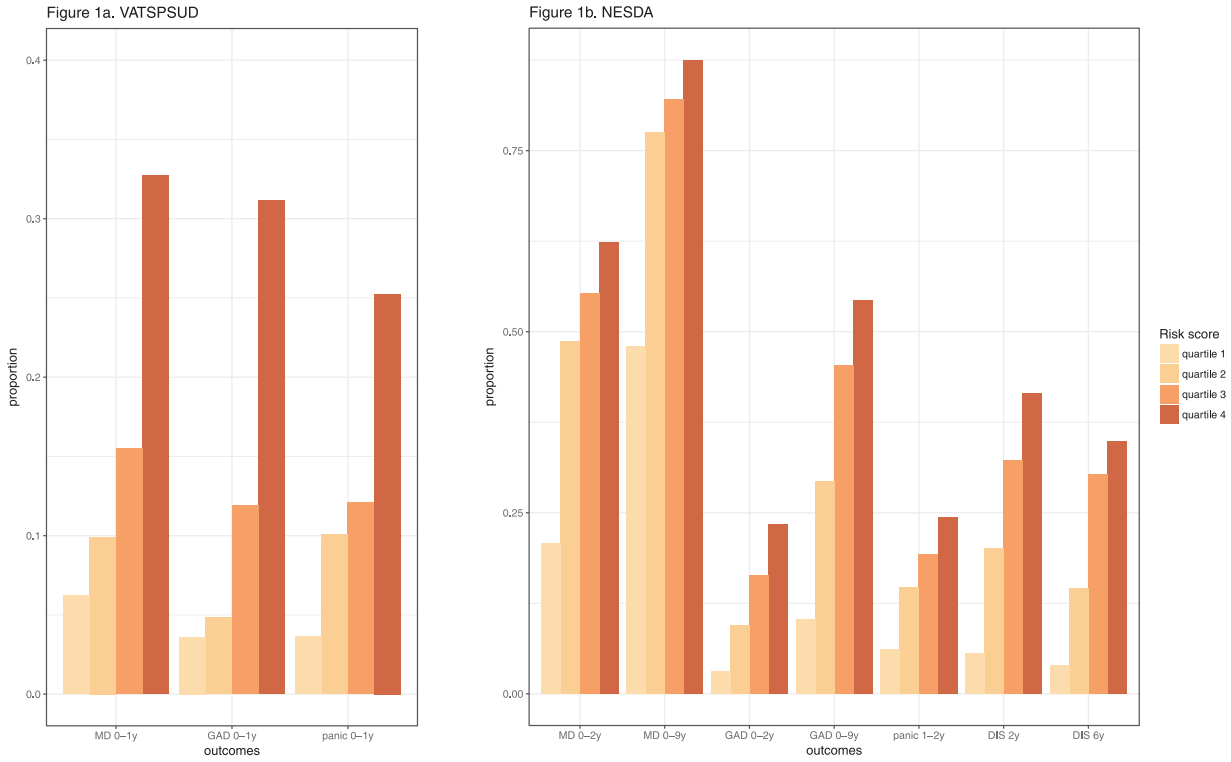

Fig. 1. Observed outcomes for different risk groups.
DIS, disability (WHODAS > 40); GAD, generalized anxiety disorder; MD, major depression; panic, panic disorder; WHODAS, World Health Organization Disability Assessment Schedule; y, year.
In each external validation sample, subjects were stratified in quartiles based on their recurrence risk score: quartile 1 includes 25% of subjects with the lowest recurrence risk scores and quartile 4 includes the 25% of subjects with the highest scores. The subjects in quartiles 2 and 3 had intermediate scores. The y-axis shows the proportion of subjects reporting the outcome during follow-up. (a) Presents the results of the VATSPSUD test sample, (b) presents the results of the NESDA test sample. The number of cases (N) with present data for each outcome are described in Table 3.
Values of unstandardized risk score in VATSPSUD: quartile 1 [0.17–0.51]; quartile 2 (0.51–0.68]; quartile 3 (0.68–0.92]; quartile 4 (0.92–2.3].
Values of unstandardized risk score in NESDA: quartile 1 [0.22–0.83]; quartile 2 (0.83–1.11]; quartile 3 (1.11–1.49]; quartile 4 (1.49–2.2].

outcomes with nearly equal performance. However, prediction was not improved when we combined the risk score and neuroticism.

4.2. Scientific and clinical relevance

First, estimates of prediction performance were similar across two differently ascertained test samples. This indicates that the combination of risk factors predicting future episodes of MD are to some extent shared rather than being unique across subjects with MD sampled from the general population, primary care, and specialized mental health care, across subjects from different countries, twins vs. non-twins, and measured with different procedures. This is a promising finding for clinical practice since a prediction model derived in one sample could be relevant for clinical populations, rather than being restricted to the training sample used to develop the model.

Second, the model predicted a broader range of adverse outcomes than it was originally developed for: it did not only predict future episodes of MD but also episodes of anxiety disorders and disability. Thus, the model could give clinicians an estimate of the risk on multiple outcomes instead of only one, which might facilitate treatment decisions. For example, one could think of decisions on the intensity of monitoring, or continuing treatment in patients who recovered from depression to prevent future episodes of MD or anxiety disorders (Coplan et al., 2015).

Third, the model predicted future episodes of MD, anxiety disorders, and disability very similarly. Partly, this was expected because of the high rates of co-occurrence between MD, anxiety disorders and severe disability, and the overlap in their risk factors (Kendler et al., 2011). However, it was surprising how similar the model predicted across these outcomes. Future studies are needed to investigate whether more specific prediction models can be identified with larger training samples.

4.3. Relation to previous studies

In previous studies using independent test data, estimates of prediction performance for models predicting course of MD were quite similar. In our study, the average AUC across future episodes of depression, anxiety, and disability was 0.71, whereas in previous studies predicting course-related outcomes in MD the AUC ranged from 0.63 to 0.76 (for ≥12 weeks follow-up) (Chekroud et al., 2016; de Vries et al., 2018; Kessler et al., 2016; Perlis, 2013; Wang et al., 2014). Interestingly, estimates of prediction performance for prediction models in other medical disciplines are not very different. For instance, similar AUC's have been found for instance in models predicting mortality after myocardial infarction (0.75–0.77) (van Loo et al., 2014b), other outcomes in cardiology (0.7–0.8) (Siontis et al., 2012), melanoma (0.7–0.8) (Usher-Smith et al., 2014), or bleeding when using antiplatelet therapy after percutaneous coronary intervention (0.64) (Yeh et al., 2016).

Despite its potential relevance for clinical practice, few previous studies assessed the predictive value of a model for a wider range of outcomes than the model was original trained for. Only one study externally validated one data mining risk score across multiple outcomes: MD persistence and chronicity, hospitalization for depression, attempted suicide, disability due to depression at time (Kessler et al., 2016). Similar to our study, this risk score predicted these different multiple outcomes (AUC's 0.63–0.76), but its predictive value for

Table 4
AUCs of the risk score compared with competing predictors.

	Sample	AUC Risk score	FH MD[1]	AAO[2]	PRS MD[3]	Neur[7]	Traumas[4]	CSA
Any MD								
MD 0–1 year[5]	VATSPSUD	0.73	0.58	0.56	n.a.	0.65	0.56	0.52
MD 0–2 year	NESDA	0.68	0.55	0.55	0.53	0.70	0.60	0.52
MD 0–9 year	NESDA	0.72	0.57	0.54	0.53	0.74	0.61	0.52
Any anxiety disorder								
GAD 0–1 year[5,6]	VATSPSUD	0.78	0.60	0.55	n.a.	0.64	0.56	0.56
GAD 0–2 year	NESDA	0.70	0.52	0.51	0.53	0.66	0.58	0.50
GAD 0–9 year	NESDA	0.73	0.53	0.53	0.55	0.73	0.61	0.52
Panic disorder 0–1 year[5]	VATPSUD	0.72	0.58	0.49	n.a.	0.76	0.55	0.61
Panic disorder 1–2 year[5]	NESDA	0.65	0.55	0.55	0.54	0.64	0.54	0.52
Disability								
WHODAS ≥40 2 year	NESDA	0.72	0.51	0.54	0.53	0.74	0.66	0.54
WHODAS ≥40 6 year	NESDA	0.73	0.56	0.52	0.51	0.72	0.64	0.53

AAO, age at onset; AUC, area under the receiving operating characteristic curve; CSA, childhood sexual abuse; MD, major depression; N, number; n.a., not available; PRS, polygenic risk score.
This table presents the AUC's of logistic regression models.
[1]Ratio of family members with MD, calculated by dividing the number of first-degree relatives with MD by the number of first-degree relatives.
[2]Age at onset of MD (years). To increase the comparability with the other predictors, we multiplied the age at onset by −1 to estimate its AUC, because a lower age at onset is associated with a higher risk of adverse outcomes.
[3]Polygenic risk score for MD; GWAS-data are not available for VATSPSUD.
[4]Traumas during lifetime in VATSPSUD; traumas during childhood in NESDA.
[5]Outcome assessed in the 12 months prior to interview wave(s).
[6]In VATSPSUD, we tested the association of the recurrence risk score with episodes of GAD with a duration of ≥1 month instead of ≥6 months in the year prior to interview, because only 4% of cases reported GAD with a duration ≥6 months, which might limit reliability of estimated associations.
[7] Because of the similarity of the AUC's for the risk score and neuroticism, 95%-confidence intervals were calculated for the AUC's of neuroticism (see Supplemental Table 5). All AUC−CI's of neuroticism were overlapping with the AUC−CI's of the risk score, except for MD 0–1 year VATSPSUD: AUC 0.65 (CI 0.61–0.68) and GAD 0–1 year VATSPSUD: AUC 0.64 (CI 0.59–0.68). Confidence intervals of the risk score are presented in Table 3.
For more details on the outcomes or competing predictors, we refer to Supplemental Table 2.

anxiety disorders was not assessed, so we cannot compare these results.

Our risk score performed only modestly better than neuroticism in predicting adverse outcomes in MD. However, the effect of neuroticism was attenuated to nonsignificant when added to the risk score in a multiple predictor model, while our risk score remained strongly predictive. Two previous studies found a similar attenuation of neuroticism's effect to predict recurrence of MD in a model including multiple predictors such as stressful life events and childhood traumas (Gopinath et al., 2007; Hardeveld et al., 2013b). One study found that neuroticism did not predict MD recurrence even in a univariate context (Hardeveld et al., 2013a). Given the inconsistent findings, future studies are warranted to investigate whether neuroticism is a consistent predictor of MD recurrence, and how it compares to our risk score.

4.4. Strengths and limitations

First, although our model's prediction performance was comparable to that of other models predicting course of MD, and other medical conditions, its performance is moderate (AUC~0.7), with relatively low positive predictive values for some of the rare outcomes. The model also needs information on 24 predictors, which may limit its value for clinical practice. Future studies are needed to assess whether the model can be improved by using larger training samples, other types of statistical learning techniques (Chekroud et al., 2016), or other types of data such as neuro-imaging, biomarkers, and molecular genetic data (Gillan and Whelan, 2017). However, the fact that our model exclusively utilizes readily available clinical information also is a strength as this reduces its associated costs and burden to patients.

Second, despite our study showed consistent prediction performance in two different independent test samples from different populations, prediction might be different for specific subgroups of patients. How well does this model predict course of MD in hospitalized patients, or in patients from different cultures? Furthermore, are results generalizable to situations in which less high-quality baseline data are available? Partly, this study showed that not all predictors need to be available or assessed with the exact same instruments –which promotes its applicability in clinical practice– but more work is needed to confirm this.

Third, not all our outcomes or predictors were optimally assessed. For instance, outcomes were assessed over the course of several years, instead of over a period of months, or decades. The latter would have provided more fine-grained information to test the risk score's prediction performance on the short and long term. More longitudinal studies are needed to collect these data. In addition, the polygenic score for MD only explains a limited percentage of the variance of MD (Wray et al., 2018), and different genetic variants might be implicated in MD onset than in MD recurrence.

Fourth, calculating the risk score by hand in clinical practice is labor intensive. We are working on a digitalized version of this prediction model, which facilitates implementing and testing this algorithm in clinical samples.

Fifth, in all probabilistic decision tools, the interpretation of probabilistic estimates is challenging. For instance, low probabilities are generally overrated, whereas high probabilities are underrated (Kahneman and Tversky, 1979). Thus, it should be carefully studied whether the application of these probabilistic decision support tools indeed improves clinical decision making in randomized controlled trials (Gillan and Whelan, 2017).

5. Conclusion

A prediction model based on 24 clinical characteristics consistently predicted multiple outcomes related to a more severe course of MD. Future studies are needed to test whether this risk prediction tool can serve as an extra source of information to differentiate high-risk from low-risk patients in clinical practice. The final aim would be to leverage the opportunities of data mining to improve insight into individual disease risk, and tailor treatment decisions to the individual patient.

Author contributions

The study was designed by HMvL and KSK. HMvL, SHA, TBB, and YM analysed the data. All authors contributed to interpretation of the results. HMvL drafted the manuscript; all other authors critically revised the manuscript, and approved its final version.

Role of the funding source

The infrastructure for the NESDA study (www.nesda.nl) is funded through the Geestkracht program of the Netherlands organization for Health Research and Development (NWO, ZonMw, grant number 10-000-1002) and financial contributions by participating universities and mental health care organizations (VU University Medical Center, GGZ inGeest, Leiden University Medical Center, Leiden University, GGZ Rivierduinen, University Medical Center Groningen, University of Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Rob Giel Onderzoekscentrum).

Further funding is provided by the Center for Medical Systems Biology (CSMB, NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), VU University's Institutes for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam, University Medical Center Groningen, Leiden University Medical Center, National Institutes of Health (NIH, R01D0042157–01A, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995). Part of the genotyping and analyses were funded by the Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health. Computing was supported by BiG Grid, the Dutch e-Science Grid, which is financially supported by NWO.

Declaration of Competing Interest

None.

Acknowledgments

We thank Dr. Charles O. Gardner for his invaluable work on the VATSPSUD data management and advice on statistical analyses.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2020.07.098.

References

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR, 4th Ed.American Psychiatric Publishing, Arlington, US.

American Psychiatric Association, 1987. Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R, 3rd rev. ed. Press Syndicate of the University of Cambridge, Cambridge.

Carpenter, J.R., Kenward, M.G., 2013. Multiple Imputation and its Application. Wiley, Chichester.

Carstensen, B., Plummer, M., Laara, E., 2017. Epi: a package for statistical analysis in epidemiology [WWW Document]. R Packag version 2.12.

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

Coplan, J.D., Aaronson, C.J., Panthangi, V., Kim, Y., 2015. Treating comorbid anxiety and depression: psychosocial and pharmacological approaches. World J. Psychiatry 5, 366–378. https://doi.org/10.5498/wjp.v5.i4.366.

Darcy, A.M., Louie, A.K., Roberts, L.W., 2016. Machine learning and the profession of medicine. JAMA 315, 551. https://doi.org/10.1001/jama.2015.18421.

de Vries, Y.A., Roest, A.M., Bos, E.H., Burgerhof, J.G.M., van Loo, H.M., de Jonge, P., 2018. Predicting antidepressant response by monitoring early improvement of individual symptoms of depression: individual patient data meta-analysis. Br. J. Psychiatry 1–7. https://doi.org/10.1192/bjp.2018.122.

Eaton, W.W., Shao, H., Nestadt, G., Lee, B.H., Bienvenu, O.J., Zandi, P., 2008. Population-based study of first onset and chronicity in major depressive disorder. Arch. Gen. Psychiatry 65, 513–520. https://doi.org/10.1001/archpsyc.65.5.513.

Frank, E., 1991. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Arch. Gen. Psychiatry 48, 851. https://doi.org/10.1001/archpsyc.1991.01810330075011.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.

Gillan, C.M., Whelan, R., 2017. What big data can do for treatment in psychiatry. Curr. Opin. Behav. Sci. 18, 34–42. https://doi.org/10.1016/j.cobeha.2017.07.003.

Gopinath, S., Katon, W.J., Russo, J.E., Ludman, E.J., 2007. Clinical factors associated with relapse in primary care patients with chronic or recurrent depression. J. Affect. Disord. 101, 57–63 S0165-0327(06)00462-9 [pii].

Hardeveld, F., Spijker, J., De Graaf, R., Hendriks, S.M., Licht, C.M., Nolen, W.A., Penninx, B.W., Beekman, A.T., 2013a. Recurrence of major depressive disorder across different treatment settings: results from the NESDA study. J. Affect. Disord. 147, 225–231. https://doi.org/10.1016/j.jad.2012.11.00810.1016/j.jad.2012.11.008.

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W.A., Beekman, A.T., 2013b. 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, 39–48. https://doi.org/10.1017/S0033291712002395.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2nd ed. Springer, New York.

Jeronimus, B.F., Kotov, R., Riese, H., Ormel, J., 2016. Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 pa. Psychol. Med. 46, 2883–2906. https://doi.org/10.1017/S0033291716001653.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Yilong, Dong, Q., Shen, H., Wang, Yongjun, 2017. Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 2, 230–243. https://doi.org/10.1136/svn-2017-000101.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291. https://doi.org/10.2307/3791683.

Kendler, K.S., Aggen, S.H., Knudsen, G.P., Roysamb, E., Neale, M.C., Reichborn-Kjennerud, T., 2011. The structure of genetic and environmental risk factors for syndromal and subsyndromal common DSM-IV axis I and all axis II disorders. Am. J. Psychiatry 168, 29–39. https://doi.org/10.1176/appi.ajp.2010.10030340.

Kendler, K.S., Gatz, M., Gardner, C.O., Pedersen, N.L., 2005. Age at onset and familial risk for major depression in a Swedish national twin sample. Psychol. Med. 35, 1573–1579. https://doi.org/10.1017/S0033291705005714.

Kendler, K.S., Prescott, C.A., 2006. Genes, Environment and Psychopathology: Understanding the Causes of Psychiatric and Substance use Disorders. Guilford Press, New York.

Kendler, K.S., Schmitt, E., Aggen, S.H., Prescott, C.A., 2008. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. Arch. Gen. Psychiatry 65, 674–682. https://doi.org/10.1001/archpsyc.65.6.674.

Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and co-morbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. Arch. Gen. Psychiatry 62, 617–627. https://doi.org/10.1001/archpsyc.62.6.617.

Kessler, R.C., van Loo, H.M., Wardenaar, K.J., Bossarte, R.M., Brenner, L.A., Cai, T., Ebert, D.D., Hwang, I., Li, J., de Jonge, P., Nierenberg, A.A., Petukhova, M.V., Rosellini, A.J., Sampson, N.A., Schoevers, R.A., Wilcox, M.A., Zaslavsky, A.M., 2016. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol. Psychiatry 21, 1366–1371. https://doi.org/10.1038/mp.2015.198.

Kraemer, H.C., 2014. Effect size. Encycl. Clin. Psychol. https://doi.org/10.1002/9781118625392.wbecp048. Major Reference Works.

Lamers, F., van Oppen, P., Comijs, H.C., Smit, J.H., Spinhoven, P., van Balkom, A.J.L.M., Nolen, W.A., Zitman, F.G., Beekman, A.T.F., Penninx, B.W.J.H., 2011. Comorbidity patterns of anxiety and depressive disorders in a large cohort study. J. Clin. Psychiatry 72, 341–348. https://doi.org/10.4088/JCP.10m06176blu.

Mangiafico, S., 2017. rcompanion: functions to Support Extension Education Program Evaluation.

Mistry, S., Harrison, J.R., Smith, D.J., Escott-Price, V., Zammit, S., 2018. The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: a systematic review. J. Affect. Disord. 234, 148–155. https://doi.org/10.1016/j.jad.2018.02.005.

Moffitt, T.E., Harrington, H., Caspi, A., Kim-Cohen, J., Goldberg, D., Gregory, A.M., Poulton, R., 2007. Depression and generalized anxiety disorder: cumulative and sequential comorbidity in a birth cohort followed prospectively to age 32 years. Arch. Gen. Psychiatry 64, 651–660 64/6/651 [pii].

Paterniti, S., Sterner, I., Caldwell, C., Bisserbe, J.-C., 2017. Childhood neglect predicts the course of major depression in a tertiary care sample: a follow-up study. BMC Psychiatry 17, 113. https://doi.org/10.1186/s12888-017-1270-x.

Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., Assendelft, W.J., Van, D.M., Verhaak, P., Wensing, M., De Graaf, R., Hoogendijk, W.J., Ormel, J., Van Dyck, R., 2008. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int. J. Methods Psychiatr. Res. 17, 121–140. https://doi.org/10.1002/mpr.256.

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

Peterson, R.E., Cai, N., Bigdeli, T.B., Li, Y., Reimers, M., Nikulova, A., Webb, B.T., Bacanu, S.A., Riley, B.P., Flint, J., Kendler, K.S., 2016. The genetic architecture of major depressive disorder in Han Chinese women. JAMA psychiatry. https://doi.org/10.

1001/jamapsychiatry.2016.3578.
R Core Team, 2017. R: a language and environment for statistical computing.
Royston, P., Altman, D.G., 2013. External validation of a Cox prognostic model: principles and methods. BMC Med. Res. Methodol. 13, 33. https://doi.org/10.1186/1471-2288-13-33.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. 39, 1–13. https://doi.org/10.18637/jss.v039.i05.
Siontis, G.C.M., Tzoulaki, I., Siontis, K.C., Ioannidis, J.P.A., 2012. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ 344, e3318. https://doi.org/10.1136/BMJ.E3318.
Usher-Smith, J.A., Emery, J., Kassianos, A.P., Walter, F.M., 2014. Risk prediction models for melanoma: a systematic review. Cancer Epidemiol. Biomark. Prev. 23, 1450–1463. https://doi.org/10.1158/1055-9965.EPI-14-0295.
van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67.
van Loo, H.M., Aggen, S.H., Gardner, C.O., Kendler, K.S., 2018. Sex similarities and differences in risk factors for recurrence of major depression. Psychol. Med. 48, 1685–1693. https://doi.org/10.1017/S0033291717003178.
van Loo, H.M., Aggen, S.H., Gardner, C.O., Kendler, K.S., 2015. Multiple risk factors predict recurrence of major depressive disorder in women. J. Affect. Disord. 180, 52–61. https://doi.org/10.1016/j.jad.2015.03.045.
van Loo, H.M., Cai, T., Gruber, M.J., Li, J., de Jonge, P., Petukhova, M., Rose, S., Sampson, N.A., Schoevers, R.A., Wardenaar, K.J., Wilcox, M.A., Al-Hamzawi, A.O., Andrade, L.H., Bromet, E.J., Bunting, B., Fayyad, J., Florescu, S.E., Gureje, O., Hu, C., Huang, Y., Levinson, D., Medina-Mora, M.E., Nakane, Y., Posada-Villa, J., Scott, K.M., Xavier, M., Zarkov, Z., Kessler, R.C., 2014a. Major depressive disorder subtypes to predict long-term course. Depress. Anxiety 31, 765–777. https://doi.org/10.1002/da.22233.
van Loo, H.M., van den Heuvel, E.R., Schoevers, R.A., Anselmino, M., Carney, R.M., Denollet, J., Doyle, F., Freedland, K.E., Grace, S.L., Hosseini, S.H., Parakh, K., Pilote, L., Rafanelli, C., Roest, A.M., Sato, H., Steeds, R.P., Kessler, R.C., de Jonge, P., 2014b. Sex dependent risk factors for mortality after myocardial infarction: individual patient data meta-analysis. BMC Med. 12, 242 s12916-014-0242-y [pii].
Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study,Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., Belbin, G., Kenny, E.E., Schierup, M.H., De Jager, P., Patsopoulos, N.A., McCarroll, S., Daly, M., Purcell, S., Chasman, D., Neale, B., Goddard, M., Visscher, P.M., Kraft, P., Patterson, N., Price, A.L., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., Pers, T.H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S.A., Begemann, M., Belliveau, R.A., Bene, J., Bergen, S.E., Bevilacqua, E., Bigdeli, T.B., Black, D.W., Bruggeman, R., Buccola, N.G., Buckner, R.L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R.M., Carr, V.J., Carrera, N., Catts, S.V., Chambert, K.D., Chan, R.C.K., Chen, R.Y.L., Chen, E.Y.H., Cheng, W., Cheung, E.F.C., Chong, S.A., Cloninger, C.R., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J.J., Curtis, D., Davidson, M., Davis, K.L., Degenhardt, F., Del Favero, J., DeLisi, L.E., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A.H., Farrell, M.S., Frank, J., Franke, L., Freedman, R., Freimer, N.B., Friedl, M., Friedman, J.I., Fromer, M., Genovese, G., Georgieva, L., Gershon, E.S., Giegling, I., Giusti-Rodrguez, P., Godard, S., Goldstein, J.I., Golimbet, V., Gopal, S., Gratten, J., Grove, J., de Haan, L., Hammer, C., Hamshere, M.L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A.M., Henskens, F.A., Herms, S., Hirschhorn, J.N., Hoffmann, P., Hofman, Andrea, Hollegaard, M.V., Hougaard, D.M., Ikeda, M., Joa, I., Julia, A., Kahn, R.S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M.C., Kelly, B.J., Kennedy, J.L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J.A., Konte, B., Kucinskas, V., Kucinskiene, Z.A., Kuzelova-Ptackova, H., Kahler, A.K., Laurent, C., Keong, J.L.C., Lee, S.H., Legge, S.E., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C.M., Lubinski, J., Lnnqvist, J., Macek, M., Magnusson, P.K.E., Maher, B.S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R.W., McDonald, C., McIntosh, A.M., Meier, S., Meijer, C.J., Melegh, B., Melle, I., Mesholam-Gately, R.I., Metspalu, A., Michie, P.T., Milani, L., Milanova, V., Mokrab, Y., Morris, D.W., Mors, O., Mortensen, P.B., Murphy, K.C., Murray, R.M., Myin-Germeys, I., Mller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D.A., Nestadt, G., Nicodemus, K.K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F.A., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Pantelis, C., Papadimitriou, G.N., Papiol, S., Parkhomenko, E., Pato, M.T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D.O., Pietilinen, O., Pimm, J., Pocklington, A.J., Powell, J., Price, A., Pulver, A.E., Purcell, S.M., Quested, D., Rasmussen, H.B., Reichenberg, A., Reimers, M.A., Richards, A.L., Roffman, J.L., Roussos, P., Ruderfer, D.M., Salomaa, V., Sanders, A.R., Schall, U., Schubert, C.R., Schulze, T.G., Schwab, S.G., Scolnick, E.M., Scott, R.J., Seidman, L.J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J.M., Sim, K., Slominsky, P., Smoller, J.W., So, H.-C., Spencer, C.C.A., Stahl, E.A., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R.E., Strengman, E., Strohmaier, J., Stroup, T.S., Subramaniam, M., Suvisaari, J., Svrakic, D.M., Szatkiewicz, J.P., Sderman, E., Thirumalai, S., Toncheva, D., Tooney, P.A., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B.T., Weiser, M., Wildenauer, D.B., Williams, N.M., Williams, S., Witt, S.H., Wolen, A.R., Wong, E.H.M., Wormley, B.K., Wu, J.Q., Xi, H.S., Zai, C.C., Zheng, X., Zimprich, F., Wray, N.R., Stefansson, K., Visscher, P.M., Adolfsson, R., Andreassen, O.A., Blackwood, D.H.R., Bramon, E., Buxbaum, J.D., Børglum, A.D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P.V., Gill, M., Gurling, H., Hultman, C.M., Iwata, N., Jablensky, A.V., Jonsson, E.G., Kendler, K.S., Kirov, G., Knight, J., Lencz, T., Levinson, D.F., Li, Q.S., Liu, J., Malhotra, A.K., McCarroll, S.A., McQuillin, A., Moran, J.L., Mortensen, P.B., Mowry, B.J., Nthen, M.M., Ophoff, R.A., Owen, M.J., Palotie, A., Pato, C.N., Petryshen, T.L., Posthuma, D., Rietschel, M., Riley, B.P., Rujescu, D., Sham, P.C., Sklar, P., St. Clair, D., Weinberger, D.R., Wendland, J.R., Werge, T., Daly, M.J., Sullivan, P.F., O'Donovan, M.C., Kraft, P., Hunter, D.J., Adank, M., Ahsan, H., Aittomäki, K., Baglietto, L., Berndt, S., Blomquist, C., Canzian, F., Chang-Claude, J., Chanock, S.J., Crisponi, L., Czene, K., Dahmen, N., Silva, I.dos S., Easton, D., Eliassen, A.H., Figueroa, J., Fletcher, O., Garcia-Closas, M., Gaudet, M.M., Gibson, L., Haiman, C.A., Hall, P., Hazra, A., Hein, R., Henderson, B.E., Hofman, Albert, Hopper, J.L., Irwanto, A., Johansson, M., Kaaks, R., Kibriya, M.G., Lichtner, P., Lindström, S., Liu, J., Lund, E., Makalic, E., Meindl, A., Meijers-Heijboer, H., Müller-Myhsok, B., Muranen, T.A., Nevanlinna, H., Peeters, P.H., Peto, J., Prentice, R.L., Rahman, N., Sánchez, M.J., Schmidt, D.F., Schmutzler, R.K., Southey, M.C., Tamimi, R., Travis, R., Turnbull, C., Uitterlinden, A.G., van der Luijt, R.B., Waisfisz, Q., Wang, Z., Whittemore, A.S., Yang, R., Zheng, W., 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. 97, 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.
Wang, J.L., Patten, S., Sareen, J., Bolton, J., Schmitz, N., MacQueen, G., 2014. Development and validation of a prediction algorithm for use by health professionals in prediction of recurrence of major depression. Depress. Anxiety 31, 451–457.
Wardenaar, K.J., van Loo, H.M., Cai, T., Fava, M., Gruber, M.J., Li, J., de Jonge, P., Nierenberg, A.A., Pethukova, M.V., Rose, S., Sampson, N.A., Schoevers, R.A., Wilcox, M.A., Alonso, J., Bromet, E.J., Bunting, B., Florescu, S.E., Fukao, A., Gureje, O., Hu, C., Huang, Y.Q., Karam, A.N., Levinson, D., Medina-Mora, M.E., Posada-Villa, J., Scott, K.M., Taib, N.I., Viana, M.C., Xavier, M., Zarkov, Z., Kessler, R.C., 2014. The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity. Psychol. Med. 44, 3289–3302.
Wickham, H., 2009. ggplot2: Elegant graphics for Data Analysis. Springer, New York.
Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A.F.T., Bigdeli, T.B., Binder, E.B., Blackwood, D.R.H., Bryois, J., Buttenschøn, H.N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J.H., Clarke, T.-K., Coleman, J.I.R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G.E., Crowley, C.A., Dashti, H.S., Davies, G., Deary, I.J., Degenhardt, F., Derks, E.M., Direk, N., Dolan, C.V., Dunn, E.C., Eley, T.C., Eriksson, N., Escott-Price, V., Kiadeh, F.H.F., Finucane, H.K., Forstner, A.J., Frank, J., Gaspar, H.A., Gill, M., Giusti-Rodríguez, P., Goes, F.S., Gordon, S.D., Grove, J., Hall, L.S., Hannon, E., Hansen, C.S., Hansen, T.F., Herms, S., Hickie, I.B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.-J., Hougaard, D.M., Hu, M., Hyde, C.L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J.A., Kohane, I.S., Kraft, J., Kretzschmar, W.W., Krogh, J., Kutalik, Z., Lane, J.M., Li, Yihan, Li, Yun, Lind, P.A., Liu, X., Lu, L., MacIntyre, D.J., MacKinnon, D.F., Maier, R.M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S.E., Mehta, D., Middeldorp, C.M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F.M., Montgomery, G.W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M.G., Nyholt, D.R., O'Reilly, P.F., Oskarsson, H., Owen, M.J., Painter, J.N., Pedersen, C.B., Pedersen, M.G., Peterson, R.E., Pettersson, E., Peyrot, W.J., Pistis, G., Posthuma, D., Purcell, S.M., Quiroz, J.A., Qvist, P., Rice, J.P., Riley, B.P., Rivera, M., Saeed Mirza, S., Saxena, R., Schoevers, R., Schulte, E.C., Shen, L., Shi, J., Shyn, S.I., Sigurdsson, E., Sinnamon, G.B.C., Smit, J.H., Smith, D.J., Stefansson, H., Steinberg, S., Stockmeier, C.A., Streit, F., Strohmaier, J., Tansey, K.E., Teismann, H., Teumer, A., Thompson, W., Thomson, P.A., Thorgeirsson, T.E., Tian, C., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A.G., Umbricht, D., Van der Auwera, S., van Hemert, A.M., Viktorin, A., Visscher, P.M., Wang, Y., Webb, B.T., Weinsheimer, S.M., Wellmann, J., Willemsen, G., Witt, S.H., Wu, Y., Xi, H.S., Yang, J., Zhang, F., Arolt, V., Baune, B.T., Berger, K., Boomsma, D.I., Cichon, S., Dannlowski, U., de Geus, E.C.J., DePaulo, J.R., Domenici, E., Domschke, K., Esko, T., Grabe, H.J., Hamilton, S.P., Hayward, C., Heath, A.C., Hinds, D.A., Kendler, K.S., Kloiber, S., Lewis, G., Li, Q.S., Lucae, S., Madden, P.F.A., Magnusson, P.K., Martin, N.G., McIntosh, A.M., Metspalu, A., Mors, O., Mortensen, P.B., Müller-Myhsok, B., Nordentoft, M., Nöthen, M.M., O'Donovan, M.C., Paciga, S.A., Pedersen, N.L., Penninx, B.W.J.H., Perlis, R.H., Porteous, D.J., Potash, J.B., Preisig, M., Rietschel, M., Schaefer, C., Schulze, T.G., Smoller, J.W., Stefansson, K., Tiemeier, H., Uher, R., Völzke, H., Weissman, M.M., Werge, T., Winslow, A.R., Lewis, C.M., Levinson, D.F., Breen, G., Børglum, A.D., Sullivan, P.F., 2018. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. 1. https://doi.org/10.1038/s41588-018-0090-3.
Xia, J., He, Q., Li, Yihan, Xie, D., Zhu, S., Chen, J., Shen, Y., Zhang, N., Wei, Y., Chen, C., Shen, J., Zhang, Yan, Gao, C., Li, Youhui, Ding, J., Shen, W., Wang, Q., Cao, M., Liu, T., Zhang, J., Duan, H., Bao, C., Ma, P., Zhou, C., Luo, Y., Zhang, F., Liu, Y., Li, Yi, Jin, G., Zhang, Yutang, Liang, W., Chen, Yunchun, Zhao, C., Li, H., Chen, Yiping, Shi, S., Kendler, K.S., Flint, J., Wang, X., 2011. The relationship between neuroticism, major depressive disorder and comorbid disorders in Chinese women. J. Affect. Disord. 135, 100–105. https://doi.org/10.1016/J.JAD.2011.06.053.
Yeh, R.W., Secemsky, E.A., Kereiakes, D.J., Normand, S.-L.T., Gershlick, A.H., Cohen, D.J., Spertus, J.A., Steg, P.G., Cutlip, D.E., Rinaldi, M.J., Camenzind, E., Wijns, W., Apruzzese, P.K., Song, Y., Massaro, J.M., Mauri, L., DAPT Study Investigators, 2016. Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 year after percutaneous coronary intervention. JAMA 315, 1735. https://doi.org/10.1001/jama.2016.3775.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.