

Original Article

Cite this article: Banerjee S *et al* (2024). Trajectories of remitted psychotic depression: identification of predictors of worsening by machine learning. *Psychological Medicine* **54**, 1142–1151. <https://doi.org/10.1017/S0033291723002945>

Received: 16 May 2023

Revised: 6 September 2023

Accepted: 12 September 2023

First published online: 11 October 2023

Keywords:


machine learning; outcome; predictors; psychotic depression; relapse; remission; residual depressive symptoms; trajectories

Corresponding author:

Alastair J. Flint;

Email: alastair.flint@uhn.ca

Trajectories of remitted psychotic depression: identification of predictors of worsening by machine learning

Samprit Banerjee¹, Yiyuan Wu¹, Kathleen S. Bingham^{2,3,4}, Patricia Marino⁵, Barnett S. Meyers⁵, Benoit H. Mulsant^{2,3}, Nicholas H. Neufeld^{2,3}, Lindsay D. Oliver³, Jonathan D. Power⁶, Anthony J. Rothschild⁷, Jo Anne Sirey⁵, Aristotle N. Voineskos^{2,3}, Ellen M. Whyte⁸, George S. Alexopoulos⁵, Alastair J. Flint^{2,4}  and on behalf of the STOP-PD II Study Group

¹Department of Population Health Sciences, Weill Cornell Medicine, New York, USA; ²Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto, Toronto, Canada; ³Centre for Addiction and Mental Health, Toronto, Canada; ⁴Centre for Mental Health, University Health Network, Toronto, Canada; ⁵Department of Psychiatry, Weill Cornell Institute of Geriatric Psychiatry, Weill Cornell Medicine, New York, USA; ⁶Department of Psychiatry, Weill Cornell Medicine, New York, USA; ⁷University of Massachusetts Chan Medical School and UMass Memorial Health Care, Worcester, USA and ⁸Department of Psychiatry, University of Pittsburgh School of Medicine and UPMC Western Psychiatric Hospital, Pittsburgh, USA

Abstract

Background. Remitted psychotic depression (MDDPsy) has heterogeneity of outcome. The study's aims were to identify subgroups of persons with remitted MDDPsy with distinct trajectories of depression severity during continuation treatment and to detect predictors of membership to the worsening trajectory.

Method. One hundred and twenty-six persons aged 18–85 years participated in a 36-week randomized placebo-controlled trial (RCT) that examined the clinical effects of continuing olanzapine once an episode of MDDPsy had remitted with sertraline plus olanzapine. Latent class mixed modeling was used to identify subgroups of participants with distinct trajectories of depression severity during the RCT. Machine learning was used to predict membership to the trajectories based on participant pre-trajectory characteristics.

Results. Seventy-one (56.3%) participants belonged to a subgroup with a stable trajectory of depression scores and 55 (43.7%) belonged to a subgroup with a worsening trajectory. A random forest model with high prediction accuracy (AUC of 0.812) found that the strongest predictors of membership to the worsening subgroup were residual depression symptoms at onset of remission, followed by anxiety score at RCT baseline and age of onset of the first lifetime depressive episode. In a logistic regression model that examined depression score at onset of remission as the only predictor variable, the AUC (0.778) was close to that of the machine learning model.

Conclusions. Residual depression at onset of remission has high accuracy in predicting membership to worsening outcome of remitted MDDPsy. Research is needed to determine how best to optimize the outcome of psychotic MDDPsy with residual symptoms.

Introduction

Major depressive disorder with psychotic features (MDDPsy) has a worse long-term outcome than non-psychotic depression, with a higher rate of relapse and recurrence, more frequent psychiatric hospitalization, and poorer long-term function (Coryell *et al.* 1996; Jääskeläinen *et al.* 2018; Nietola *et al.* 2018). There is, however, heterogeneity of outcome, with some individuals progressing to full recovery, while others have a brittle or relapsing course despite adequate treatment.

To our knowledge, no study has examined trajectories of outcome of remitted MDDPsy. With respect to non-psychotic depression, Gueorguieva, Chekroud, and Krystal (2017) performed a post hoc analysis of data from double-blind discontinuation trials of fluoxetine or duloxetine *v.* placebo among individuals who had responded to acute treatment of MDD and identified a 'relapse' trajectory and two trajectories of stable depression scores. Female sex, shorter length of time with clinical response, and higher residual depression severity at discontinuation baseline increased the odds of belonging to the relapse trajectory.

STOP-PD II was a randomized clinical trial that examined the clinical outcomes of persons aged 18–85 years who had experienced 8 weeks of sustained remission or near-remission of MDDPsy when treated with sertraline plus olanzapine and were then randomized to 36

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

weeks of sertraline plus olanzapine or sertraline plus placebo (NCT01427608) (Flint *et al.* 2019). As hypothesized, participants randomized to sertraline plus placebo had a higher risk of relapse than those randomized to sertraline plus olanzapine. Nevertheless, 20% of individuals in the olanzapine group experienced a relapse, while 45% in the placebo group did not, indicating significant heterogeneity of outcome in relation to treatment assignment.

In order to better understand the heterogeneity of outcome of remitted psychotic depression, and factors that contribute, we analyzed data from STOP-PD II and had two aims. First, we sought to identify subgroups of participants with distinct trajectories of depressive symptoms during the randomized phase of STOP-PD II. Second, we used machine learning to detect characteristics of participants that predicted membership to the worsening trajectory of depressive symptoms during the randomized phase, regardless of assigned treatment. The identification of predictors of individuals with the worse outcome has the potential to inform personalized care. We used a machine learning approach because of its sensitivity and replicability, ability to detect complex non-linear patterns among predictors, and ability to examine many variables, even correlated ones, simultaneously (Chekroud *et al.* 2021).

Methods

Participants and study design

The design and methods of STOP-PD II have been previously described (Flint *et al.* 2013). The study was conducted at four medical centers (University Health Network, Toronto; University of Massachusetts Chan Medical School; University of Pittsburgh School of Medicine; and Weill Cornell Medicine) following approval by their Institutional Review Boards. Written informed consent was obtained from all participants or their substitute decision maker prior to the initiation of any research procedures.

The study had 3 phases: an acute phase lasting up to 12 weeks, an 8-week stabilization phase, and a 36-week randomized controlled trial (RCT). At the time of enrollment in the acute phase of the study, participants were aged between 18 and 85 years, met Structured Clinical Interview for DSM-IV-TR (SCID) (First, Spitzer, Gibbon, & Williams, 2001) criteria for a current major depressive episode with at least one associated delusion (with or without hallucinations), and had a 17-item Hamilton Depression Rating Scale (Ham-D₁₇) (Hamilton, 1960) total score ≥ 21 . The study's exclusion criteria included DSM-IV-TR criteria for: lifetime bipolar disorder, any other psychotic disorder, or intellectual disability; current body dysmorphic disorder or obsessive-compulsive disorder; substance abuse or dependence within the preceding 3 months; and dementia preceding the index episode of depression or a 26-item IQCODE (Jorm, 2004) mean score ≥ 4 at acute phase baseline. Additional exclusions were type 1 diabetes mellitus; neurologic disease that might affect neuromuscular function; and unstable physical illness, although many of the study participants had stable chronic physical problems.

In the acute phase, participants received a combination of open-label sertraline (target dosage of 150–200 mg/day) plus open-label olanzapine (target dosage of 15–20 mg/day). The only other psychotropic medications allowed were 'as needed' lorazepam to a maximum dosage of 3 mg/day or 'as needed'

benztropine to a maximum dosage of 2 mg/day. Participants entered the stabilization phase as soon as they met the study's criteria for remission or, failing that, met criteria for near-remission at Week 12 of the acute phase. Remission was defined as the absence of delusions and hallucinations and a Ham-D₁₇ total score ≤ 10 for two consecutive weeks. Near-remission was defined as the absence of delusions and hallucinations, a Ham-D₁₇ score of 11–15 with $\geq 50\%$ reduction in baseline Ham-D₁₇ score, and rated as 'very much improved' or 'much improved' on the Clinical Global Impression (CGI) Scale (Guy, 1976). At the end of the 8-week stabilization phase, participants who still met remission or near-remission criteria following open-label treatment with sertraline plus olanzapine, and had a Mini-Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975) score ≥ 24 , were eligible for the RCT.

All participants continued to take open-label sertraline for the duration of the 36-week RCT. They were randomized under double-blind conditions to either continue olanzapine or switch from olanzapine to identically appearing placebo pills during a protocolized 4-week taper of olanzapine. Participants in the RCT were assessed weekly for the first 8 weeks and once every 4 weeks after that until they reached one of the three study end points: relapse (see Flint *et al.* 2019 for relapse criteria), study completion at RCT Week 36, or early termination.

Outcome

The outcome for the current analysis was depression severity, measured with the GRID version of the Ham-D₁₇ (Williams *et al.* 2008) at each assessment point of the RCT. There were up to 15 post-baseline assessments. (Psychosis was not considered for outcome, since sustained absence of delusions and hallucinations was an eligibility requirement for the RCT and re-emergence of psychosis during the RCT was, by virtue of being one of the criteria of relapse (Flint *et al.* 2019), a study end point).

Predictor variables and associated measures

Potential predictors of depression severity trajectory membership were selected based on their previously reported association with outcome of MDD following acute treatment (Alexopoulos *et al.* 2000; Buckman *et al.* 2018; Burcusa & Iacono, 2007; Hardeveld, Spijker, De Graaf, Nolen, & Beekman, 2010; Klein, Holtman, Bockting, Heymans, & Burger, 2018; Wojnarowski, Firth, Finegan, & Delgado, 2019). In addition, the following variables were also selected as possible predictors: sociodemographic variables; study site; dosages of sertraline and olanzapine at RCT baseline; and acute and stabilization phase measures of medication-associated parkinsonism and akathisia, given their overlap with psychomotor disturbance which has been associated with risk of relapse of MDD (Flint *et al.* 2021). Table 1 lists all variables. Online Supplementary Table 1 reports the time schedule of measurement of predictor variables.

Of the clinical variables, depression severity was measured with the GRID version of the Ham-D₁₇ (Williams *et al.* 2008); delusion severity was measured with the delusion severity item of the Schedule for Affective Disorders and Schizophrenia (SADS) (Spitzer & Endicott, 1979); anxiety severity was measured with the anxiety subscale of the Hospital Anxiety and Depression Scale (HADS-A) (Zigmond & Snaith, 1983); severity of psychomotor disturbance was measured with the CORE instrument (Parker *et al.* 1993); clinical global impression was measured

Table 1. Characteristics of subgroups based on latent growth mixture model trajectories of Hamilton Depression Rating Scale total scores during the STOP-PD II randomized controlled trial ($N = 126$)

	N missing	Worsening trajectory ($N = 55$)	Stable trajectory ($N = 71$)	Test statistic	df	P
Characteristics at Acute Phase Baseline						
Age, Mean (s.d.), (years)	0	55.6 (15.7)	55.1 (14.4)	$t = 0.17$	111	0.86
Gender, N (%)	0			$\chi^2 = 0.12$	1	0.72
Men		20 (36.4)	28 (39.4)			
Women		35 (63.6)	43 (60.6)			
Race, N (%)	0			$\chi^2 = 1.38$	2	0.50
White		47 (85.4)	56 (78.9)			
Black		6 (10.9)	9 (12.7)			
Other ^a		2 (3.6)	6 (8.4)			
Hispanic ethnicity, N (%)	0	5 (9.1)	10 (14.1)	$\chi^2 = 0.74$	1	0.39
Marital, N (%)	0			$\chi^2 = 0.47$	3	0.93
Single		16 (29.1)	18 (25.4)			
Married		26 (47.3)	37 (52.1)			
Separated/Divorced		9 (16.4)	10 (14.1)			
Widowed		4 (7.3)	6 (8.5)			
English first language, N (%)	0	48 (87.3)	51 (71.8)	$\chi^2 = 4.39$	1	0.04
Education, Mean (s.d.), (years)	0	14.0 (3.2)	13.9 (3.9)	$t = 0.09$	124	0.93
Living arrangements, N (%)	0			$\chi^2 = 0.12$	1	0.73
Lives alone		11 (20.0)	16 (22.5)			
Lives with others		44 (80.0)	55 (77.5)			
Inpatient status at acute phase enrollment, N (%)	0	36 (65.5)	51 (71.8)	$\chi^2 = 0.59$	1	0.44
Study site, N(%)	0			$\chi^2 = 7.33$	3	0.06
Cornell		10 (18.2)	21 (29.6)			
U Mass		18 (32.7)	11 (15.5)			
Pittsburgh		10 (18.2)	9 (12.7)			
Toronto		17 (30.9)	30 (42.3)			
Number of lifetime depressive episodes, N (%)	0			$\chi^2 = 6.17$	2	<0.05
1		8 (14.5)	24 (33.8)			
2–3		42 (76.4)	41 (57.5)			
>3		5 (9.1)	6 (8.5)			
Duration of current episode of depression, Median (IQR), (months)	2	6 (2,16)	5 (2,11.5)	$W = 2049$	-	0.40
Age of onset of first major depressive episode, Median (IQR), (years)	3	30 (17.25,45)	40 (28,52)	$W = 1272$	-	0.003
Treatment resistance in current episode, N (%) ^b	0	5 (9.1)	3 (4.2)	$\chi^2 = 1.23$	1	0.27
Suicide attempt in current episode, N (%)	0	12 (21.8)	13 (18.3)	$\chi^2 = 0.24$	1	0.62
Ham-D ₁₇ total score, Mean (s.d.)	0	29.1 (4.3)	28.3 (4.8)	$t = 1.0$	121	0.31
SADS delusion score, Median (IQR)	0	5 (5,6)	6 (5,6)	$W = 1592$	-	0.06
HADS anxiety score, Mean (s.d.)	5	12.8 (3.6)	12.1 (4.2)	$t = 0.9$	117	0.37
CORE total score, Median (IQR)	0	11 (5,16)	11 (5,16)	$W = 1927$	-	0.90
BPRS, Mean (s.d.)	0	51.6 (8.2)	50.6 (8.7)	$t = 0.63$	119	0.53
CGI severity, Median (IQR)	0	5 (5,5)	5 (5,6)	$W = 1736$	-	0.25

(Continued)

Table 1. (Continued.)

	N missing	Worsening trajectory (N = 55)	Stable trajectory (N = 71)	Test statistic	df	P
CIRS-G total score, Median (IQR)	0	4 (2,7)	3 (0.5, 5)	W = 2392	-	0.03
Simpson Angus Scale Total score, Median (IQR)	1	2 (0,4)	1 (0,3)	W = 2021	-	0.62
Barnes Akathisia Rating Scale global score, N (%)	0			$\chi^2 = 4.48$	2	0.11
0		52 (94.5)	59 (83.1)			
1		3 (5.5)	9 (12.7)			
2		0 (0)	3 (4.2)			
Characteristics at Acute Phase Termination						
Ham-D ₁₇ total score, Mean (s.d.)	0	7.9 (3.2)	4.6 (2.8)	t = 6.0	105	<0.001
Characteristics at RCT Baseline						
HADS anxiety score, Mean (s.d.)	1	6.5 (4.0)	3.8 (3.2)	t = 4.1	102	<0.001
CORE total score, Median (IQR)	0	2 (1,6)	0 (0,2.5)	W = 2662	-	<0.001
CIRS-G total score, Median (IQR)	0	4 (2,7)	3 (0,4)	W = 2444	-	<0.002
CGI severity, Median (IQR)	0	1 (1,2)	1 (1,1)	W = 2496	-	<0.001
MMSE, Mean (s.d.)	0	28.0 (1.8)	28.0 (2.1)	t = 0.18	123	0.86
DKEFS Color-Word Interference Test, Mean (s.d.) ^c	8	8.0 (2.7)	8.2 (3.1)	t = -0.039	114	0.70
RBANS Coding, Mean (s.d.)	6	4.8 (3.5)	5.4 (4.0)	t = -0.081	114	0.42
RBANS List Recall, Mean (s.d.)	6	62.8 (26.3)	64.6 (25.0)	t = -0.039	107	0.70
Simpson Angus Scale Total score, Median (IQR)	0	1 (0,3)	0 (0,1)	W = 2324	-	0.053
Barnes Akathisia Rating Scale global score, N (%)	0			Fisher's exact	-	0.014
0		50 (90.9)	71 (100)			
1		4 (7.3)	0 (0)			
2		1 (1.8)	0 (0)			
Sertraline dosage, N (%), (mg/day)	0			$\chi^2 = 4.19$	2	0.12
100		4 (7.3)	9 (12.7)			
150		21 (38.2)	36 (50.7)			
200		30 (55.5)	26 (36.6)			
Olanzapine dosage, N (%), (mg/day)	0			$\chi^2 = 7.7$	3	0.054
5		1 (1.8)	8 (11.3)			
10		9 (16.4)	18 (25.4)			
15		22 (40.0)	27 (38.0)			
20		23 (41.8)	18 (25.4)			

Abbreviations: BPRS, Brief Psychiatric Rating Scale; CGI, Clinical Global Impression; CIRS-G, Cumulative Illness Rating Scale for Geriatrics; CORE, the CORE measure of psychomotor disturbance; DKEFS, Delis-Kaplan Executive Function System; HADS, Hospital Anxiety and Depression Scale; HAM-D₁₇, 17-item Hamilton Depression Rating Scale; MMSE, Mini Mental State Examination; RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; SADS, Schedule for Affective Disorders and Schizophrenia.

^a'Other Race' includes American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and unknown or not reported.

^bTreatment resistance defined as an antidepressant plus antipsychotic combination rating score of 3 or higher on the Antidepressant Treatment History Form or seven or more treatments of electroconvulsive therapy during the current episode of psychotic depression.

^cDKEFS color-word interference test condition 3 final weighted score.

with the Clinical Global Impression Scale severity item (Guy, 1976); overall severity of illness at acute baseline was measured with the Brief Psychiatric Rating Scale (Overall & Gorham, 1962); lifetime medical burden was quantified by the Cumulative Illness Rating Scale for Geriatrics (CIRS-G) (Miller *et al.* 1992); treatment resistance during the index episode of psychotic depression was defined on the Antidepressant

Treatment History Form (ATHF) (Oquendo *et al.* 2003) as an antidepressant plus antipsychotic combination rating score of 3 or higher or seven or more treatments of electroconvulsive therapy (Blumberger *et al.* 2011); global cognitive function was measured with the MMSE (Folstein *et al.* 1975); executive function was measured with the Delis-Kaplan Executive Function System (DKEFS) (Delis, Kaplan, & Kramer, 2001) Color Word

Interference Condition 3 final weighted scaled score (a continuous measure of inhibition); information processing speed was measured with the Coding task from the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) (Randolph, 1998); delayed verbal recall was measured with the RBANS List Recall task; medication-associated parkinsonism was measured with the Simpson Angus Scale (Simpson & Angus, 1970); and akathisia was measured with the Barnes Akathisia Rating Scale (Barnes, 1989). Age of onset of the first lifetime episode of MDD, lifetime number of episodes of MDD, duration of the index episode of depression, and presence or absence of a suicide attempt during the index episode of depression were assessed at acute baseline using the SCID.

Data analyses

Subgroups based on trajectories of depression scores during the RCT

Latent class mixed modeling (LCMM) was used to identify subgroups of participants with distinct longitudinal trajectories of depression severity during the 36-week RCT. A range of latent subgroups ($K=1$ to 4) were considered. For each K -subgroup model, various shapes of the depression trajectory were considered i.e. constant, linear, or non-linear (quadratic or cubic), and the final shape was chosen based on model fit statistics (Bayesian information criterion or BIC). Average posterior probabilities of group membership were used as a measure of internal reliability for each trajectory. The posterior probability of membership is computed using the Bayes rule and denotes the probability of an individual belonging to a certain trajectory class conditional on the individuals' repeated measures of the outcome. Posterior probability values greater than 0.70–0.80 for each trajectory subgroup suggest greater homogeneity within a trajectory group than between trajectory groups.

Predictors of membership of subgroups based on trajectories of depression score

To identify membership to the trajectories of depression severity during the RCT, a number of sociodemographic and clinical characteristics of participants were considered (Table 1). Specifically, three sets of predictors were considered: (1) predictors measured at baseline of the acute phase; (2) predictors measured at the RCT baseline (end of the stabilization phase); and (3) longitudinal features of predictor variables during the acute and stabilization phases. Severity of depression at RCT baseline was not included as a predictor because it was included in deriving the trajectories of depression severity; instead, we used Ham-D₁₇ total score at the end of the acute phase, when participants first met criteria for remission/near-remission (hereafter referred to as 'remission') of psychotic depression. The third group of predictors extracted longitudinal features from variables that had repeated measures during the acute and stabilization phases of the study, specifically delusion severity, clinical global impression scale severity, and medication-associated parkinsonism and akathisia. The longitudinal features considered were median, standard deviation (s.d.), change from acute baseline to RCT baseline, a binary indicator of monotonic increasing or decreasing trend during the acute and stabilization phase, and a Spearman's correlation with time. Any predictors at acute baseline or RCT baseline that had more than 30% missing values were not used in this analysis. Missing values of remaining predictors were imputed using the proximity measures of a random forest (Stekhoven & Buehlmann, 2012).

Initially, acute baseline and RCT baseline characteristics of the LCMM-identified latent trajectories were compared using chi-square or independent two-sample t tests as appropriate. Then, a machine learning algorithm, random forest, was used to predict membership to the groups of trajectories using the predictors mentioned above. Briefly, the random forest algorithm considers complex interactions between predictors using decision trees and averages predictions (by majority voting) over multiple decision trees in bootstrapped samples. Predictors were ranked on their relative 'importance' by quantifying the improvement in prediction error by each predictor using the Gini Impurity Index. We report the set of predictors that explains 70% of cumulative reduction of Gini Index.

Prediction accuracy was operationalized using the area under the receiver operating characteristics curve (AU-ROC) and was estimated by five-fold cross-validation (CV). The variability of the estimated AU-ROC is reported by computing the 95% bootstrapped confidence intervals of the AUC. Tuning parameters for random forest (number of trees and the number of randomly chosen predictors for a candidate split in each tree) were identified by embedding another 5-fold CV within the outer CV.

Results

Of the 269 participants enrolled in the acute phase of STOP-PD II, 126 participated in the RCT. The CONSORT figure has been reported elsewhere (Flint *et al.* 2019).

Trajectories of depression scores during the RCT

A two-group trajectory model was chosen based on BIC, average posterior probability of group membership, sample size in each group, and clinical interpretability. This model was characterized by a stable trajectory ($n=71$; 56.3% of participants) and a worsening trajectory ($n=55$; 43.7% of participants) (Fig. 1). The average posterior probabilities for membership in the stable and worsening trajectories were 0.99 and 0.98, respectively, showing strong internal reliability of each trajectory group; that is, participants within a trajectory group were more homogeneous than between the two groups. The depression severity of the worsening trajectory had a linear increasing trend over the RCT (estimate of slope = 0.7765, 95% Bootstrap CI 0.5357–1.0175), while that of the stable trajectory did not change over time (estimate of slope = 0.1461, 95% Bootstrap CI –0.0291 to 0.3212). The group with the worsening trajectory had significantly higher mean (s.d.) depression severity at RCT baseline compared to the stable trajectory group (7.82 [3.31] *v.* 3.68 [2.60] respectively; $t=7.6$, $df=100$, $p<0.001$). Rates of relapse were 70.9% (39/55) and 11.3% (8/71) in the worsening and stable trajectory groups, respectively.

Sixty one percent (43/71) of participants with a stable trajectory were in the sertraline plus olanzapine randomized group and 39% (28/71) were in the sertraline plus placebo group (mean [s.d.] Ham-D₁₇ total score at onset of remission = 4.72 [2.82] and 4.36 [2.72], respectively). Of the participants with a worsening trajectory, 38% (21/55) were in the sertraline plus olanzapine randomized group and 62% (34/55) were in the sertraline plus placebo group (mean [s.d.] Ham-D₁₇ total score at onset of remission = 8.95 [3.38] and 7.21 [2.96], respectively).

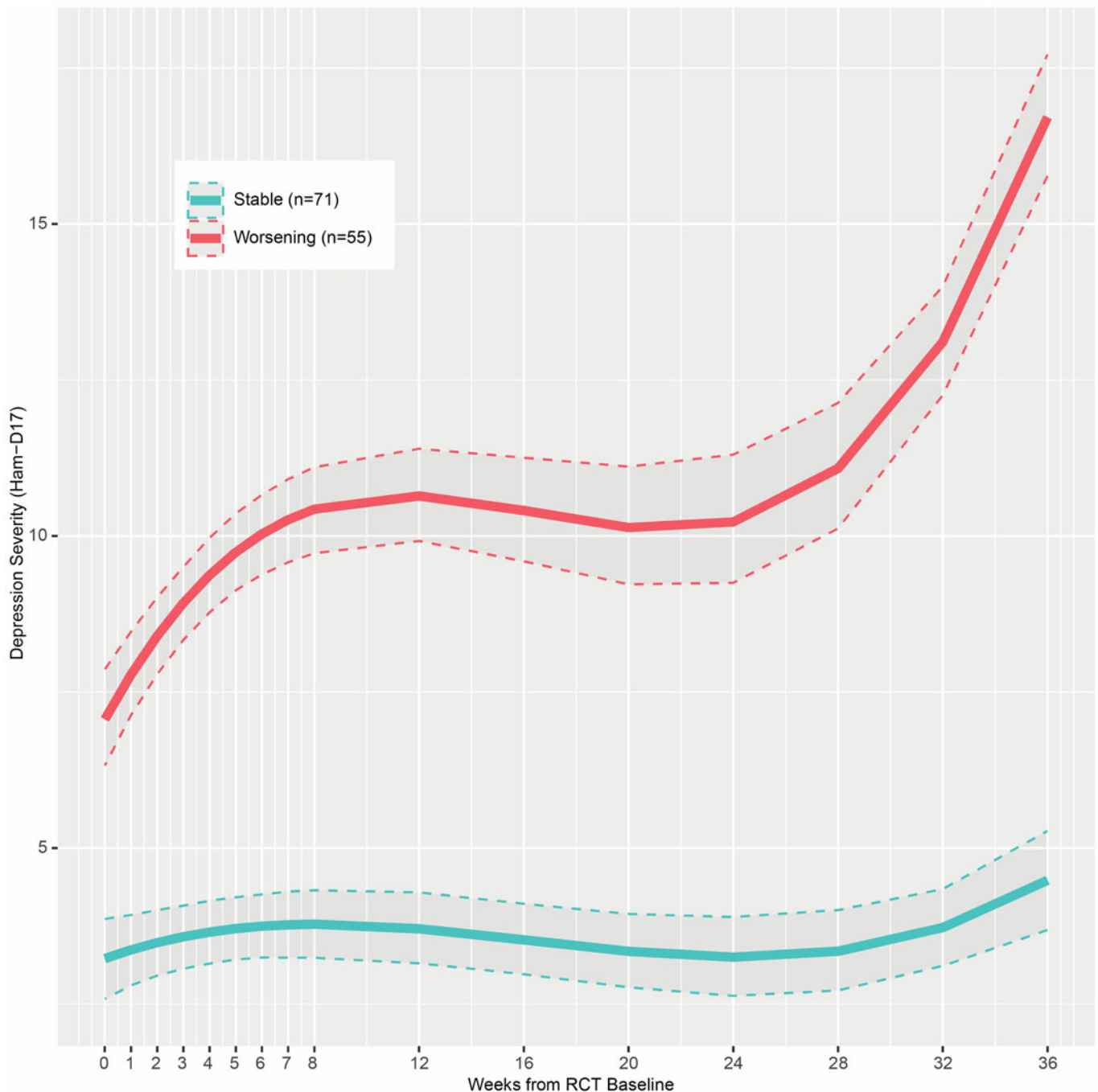


Figure 1. Latent Growth Mixture Model of estimated trajectories of depression severity (along with 95% bootstrapped confidence intervals) among participants in the randomized phase of STOP-PD II ($N = 126$).

Predictors of worsening depression trajectory: machine learning model

At the acute phase baseline (when participants were depressed), individuals with a worsening trajectory had earlier age of onset of the first lifetime episode of MDD, greater number of lifetime depressive episodes, greater medical burden, and were more likely to speak English as their first language (Table 1). With respect to variables assessed at RCT baseline (when participants were in remission), the worsening trajectory group had higher scores on measures of anxiety, psychomotor disturbance, clinical global

impression, medical burden, and akathisia (Table 1). Participants in the worsening trajectory group had a higher mean (s.d.) Ham-D₁₇ total score at onset of remission (i.e. at acute phase termination) than those with a stable trajectory (Table 1).

The random forest model had an AUC of 0.812 (95% CI 0.658–0.938) in predicting the worsening depression trajectory. Based on the mean decrease in Gini Impurity Index, the strongest predictors of worsening trajectory were depression score at onset of remission, followed by anxiety score at RCT baseline, and age of onset of first lifetime depressive episode (Fig. 2). Figure 2 shows

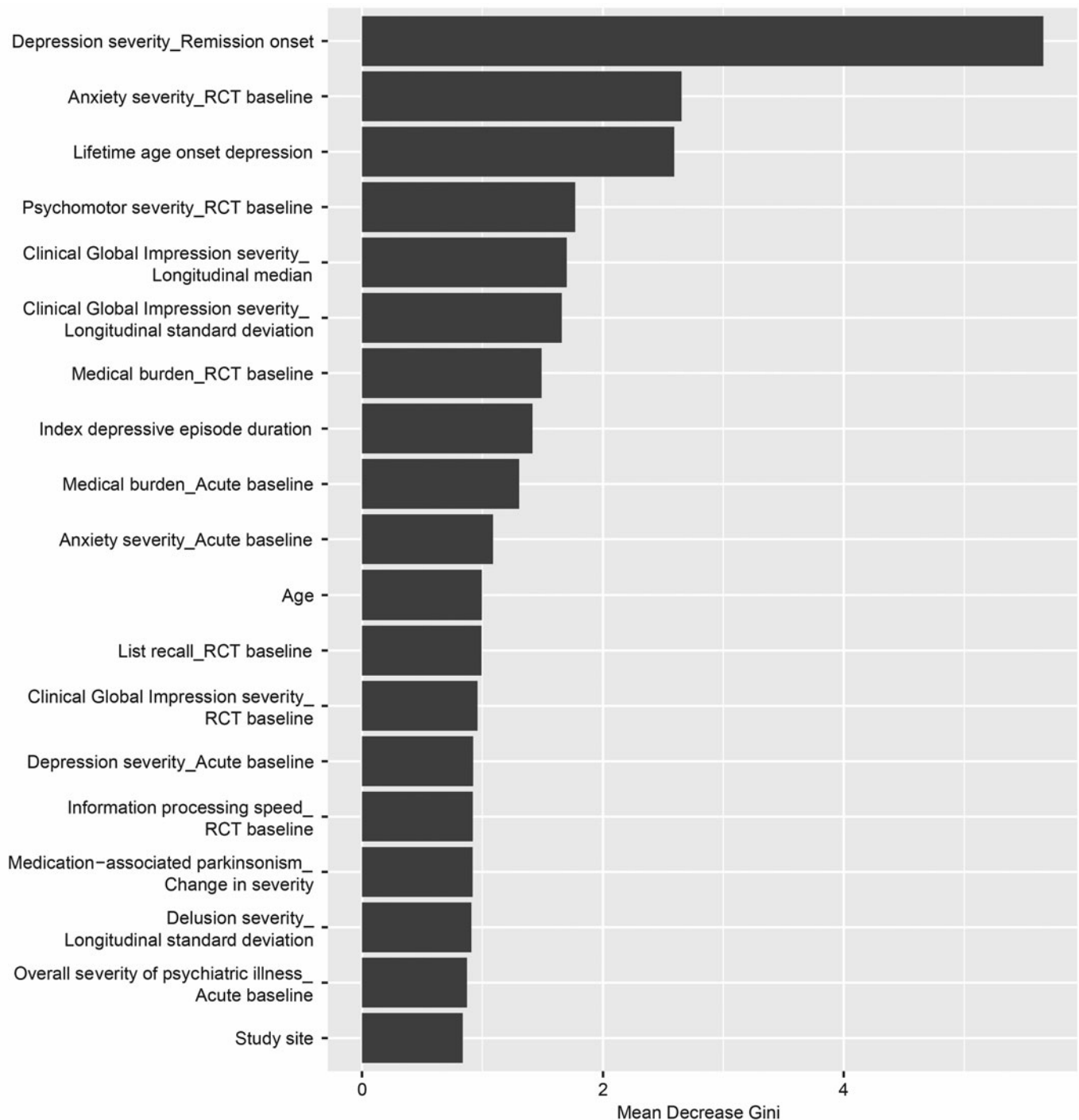


Figure 2. Variable importance in predicting membership of the worsening depression trajectory subgroup among participants in the randomized phase of STOP-PD II ($N = 126$). Predictors are presented from top to bottom in order of importance. (The horizontal axis represents mean decrease in Gini Impurity Index, which is a weighted average of reduction in leaf node impurities).

the relative importance of the set of predictors that cumulatively explain 70% reduction in the Gini Impurity Index.

We conducted two secondary analyses to further examine the relative contribution of depression severity at onset of remission as a predictor of worsening trajectory. First, we determined the predictive accuracy of the random forest model *without* depression severity at onset of remission. This revised model had an AUC of 0.772 (95% CI 0.607–0.906): anxiety severity at RCT

baseline, followed by age of onset of first lifetime depressive episode, and the median of CGI scores through the acute and stabilization phases were the strongest predictors of outcome (online Supplementary Figure 1). Second, we conducted a logistic regression model where depression severity at onset of remission was the *only* predictor variable, to determine how the prediction accuracy of a simple regression model compared with that of the machine learning model. The AUC of the logistic regression

model, estimated with a 5-fold cross-validation (same as the random forest model), was 0.778 (95% CI 0.579–0.939).

Discussion

We identified two groups of persons with remitted MDDPsy based on trajectories of depression scores during continuation treatment with either sertraline plus olanzapine or sertraline plus placebo: those with a stable trajectory and those with a worsening trajectory. In the random forest model, the strongest predictor of membership in the worsening group was the severity of depression at onset of remission. The prediction accuracy of the machine learning model that included multiple predictor variables, including depression severity, was only marginally better than that of a logistic regression model that included depression severity alone (AUCs of 0.812 and 0.778, respectively). In contrast, a secondary random forest model that did not include depression severity was slightly less accurate (AUC of 0.772) than depression severity alone.

These findings have clinical implications. First, they indicate that assessing the severity of MDDPsy in remission is important in predicting outcome. The mean (s.d.) Ham-D₁₇ score at the onset of remission was 4.6 (2.8) in the stable group and 7.9 (3.2) in the worsening group. Although both mean scores fall within the conventional range of remission of MDD in mid-life and older adults (Frank *et al.* 1991; Reynolds *et al.* 2006), our findings indicate that there is value in knowing where an individual lies within this range, since lower scores predict a better outcome.

Second, measuring severity of depression in remission may be sufficient to predict clinical course during continuation treatment, without the need to assess other prognostic variables. Although the random forest model that contained multiple predictors was slightly more accurate than depression score alone, the 3.4% difference in their AUC is of doubtful clinical significance. This may indicate that depression severity at remission is correlated with other predictors considered in the random forest model. Interestingly, Song *et al.* (2023) found that subsyndromal depression severity predicted future case-level depression with similar accuracy to that of a machine learning model that included a number of psychological and non-psychological variables (AUCs of 0.764 and 0.791, respectively). From the perspective of patient and clinician, the administration of one measure (depression severity) is less burdensome, and more likely to lead to uptake in clinical practice, than the need to administer several measures.

Third, our findings raise an important question about state *v.* trait effects in the outcome of remitted MDDPsy. In some individuals, it is possible that a more vigorous and/or a different treatment could further lessen depression severity, or eliminate symptoms entirely, thus resulting in a stable outcome (state effect). However, other individuals may be unable to achieve an asymptomatic state regardless of what treatment they receive and remain at risk for an unfavorable course of depression. In this case, the inability to achieve full remission serves as a trait marker of a phenotype that may have a distinct neurobiological signature. Further research is needed to disentangle these effects, and identify individuals with residual depression who can benefit from more vigorous or different treatment and determine how they differ biologically from persons with a trait-determined course.

Anxiety score at randomization baseline, when participants were in remission, also emerged as a strong predictor of worse outcome trajectory. The association of anxiety, either as a residual

symptom of depression or as a comorbid disorder, with worse long-term outcome of depression is well documented (Buckman *et al.* 2018; Hardeveld *et al.* 2010). In addition, psychomotor score at randomization baseline was a leading predictor of worse outcome. Although several factors have been implicated in risk of relapse and recurrence of MDD (Buckman *et al.* 2018; Hardeveld *et al.* 2010; Wojnarowski *et al.* 2019), our collective findings suggest that a focus on residual symptoms may be a fruitful area of inquiry to understand why some people with remitted MDDPsy have a poorer outcome than others.

There are limitations to the study. First, analyses were post hoc. Second, history of childhood trauma, personality traits, quality of interpersonal relationships, and life events and difficulties, which have been found to be relevant to relapse and recurrence of MDD (Buckman *et al.* 2018; Burcusa & Iacono, 2007; Perlman *et al.* 2019), were not assessed in STOP-PD II and therefore could not be included in the machine learning model. Third, the study design precluded a nuanced analysis of psychosis in the machine learning model. In order to be eligible for the RCT, participants had to experience full remission of delusions and hallucinations during acute treatment and, then, remain free of psychosis during the 8-week stabilization phase. Thus, in contrast to depression severity, there was minimal variability in SADS delusion and hallucination scores at the point of remission and during stabilization. Fourth, the trajectory of Ham-D scores in the RCT was influenced by randomized treatment, in that participants treated with sertraline plus placebo had greater risk of relapse (which would have influenced the trajectory of Ham-D scores) than participants treated with sertraline plus olanzapine. It is possible that different variables would have predicted membership of worsening trajectory if the sample had been restricted to persons treated with sertraline plus olanzapine. However, the sample size was not large enough to allow separate analyses of each treatment group. Fifth, our findings are limited to the drugs that were used in STOP-PD II. Trajectories and/or predictors may have been different had other drug combinations been used.

Strengths of the study include the well characterized sample, the rigorous approach to measurement of variables, and the large and diverse number of variables available for the machine learning model (notwithstanding the absence of the specific variables noted above). Other strengths are the wide age range of participants in keeping with the broad age of onset of MDDPsy, the 36-week duration of the RCT, and the multiple points of measurement of the outcome variable.

To conclude, the strongest predictor of membership in the worsening subgroup was the severity of depression at onset of remission. The predictive accuracy of remission depression severity alone was comparable to that of a machine learning model that included multiple variables. This finding suggests that assessing the severity of depression at onset of remission may be sufficient to inform the outcome of depression during continuation treatment, without the need to measure other prognostic variables. Finally, there is a need for research to determine how to optimize the outcome of treated psychotic depression when residual symptoms persist.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S0033291723002945>.

Acknowledgements. We thank the members of the STOP-PD II Study Group for their contributions. Members of the STOP-PD II Study Group were: Peter Giacobbe, M.D. and Brenda Swampillai B.Sc. at the University

Health Network, Toronto, and James Kennedy, M.D. and Bruce Pollock M.D., Ph.D. at the Centre for Addiction and Mental Health, Toronto; Kristina Deligiannidis, M.D., Chelsea Kosma, M.A., Wendy Marsh, M.D. at the University of Massachusetts Chan Medical School and UMass Memorial Health Care, Worcester, MA; Ariel Gildengers M.D., Joelle Kinman Ph.D., Meryl Butters Ph.D., and Michelle Zmuda M.A. at the University of Pittsburgh School of Medicine, Pittsburgh, PA; and Judith English, M.A., James Kocsis, M.D., Barbara Ladenheim, Ph.D., Vassilios Latoussakis, M.D., and Nikhil Palekar, M.D. at Weill Cornell Medicine and New York Presbyterian Hospital, NY.

Funding statement. This study was funded by USPHS grants MH 62446, MH 62518, MH 62565, and MH 62624 from the National Institute of Mental Health (NIMH). Eli Lilly provided olanzapine and matching placebo pills and Pfizer provided sertraline; neither company provided funding for the study.

The National Institute of Mental Health (NIMH) participated in the implementation of this study through the U01 mechanism. NIMH did not participate in the design of the study or the collection, management, or analysis of data. A data safety monitoring board at NIMH provided data and safety monitoring. Neither Eli Lilly nor Pfizer participated in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Competing interests. S. Banerjee has no disclosures. Y.Wu has no disclosures. K.S. Bingham receives grant support from the University of Toronto. P. Marino received research support from the NIMH at the time this work was done. B.S. Meyers received research support from the NIMH at the time this work was done. B.H. Mulsant holds and receives support from the Labatt Family Chair in Biology of Depression in Late-Life Adults at the University of Toronto. He currently receives or has received research support during the past three years from Brain Canada, the Canadian Institutes of Health Research, the CAMH Foundation, the Patient-Centered Outcomes Research Institute (PCORI), the US National Institute of Health (NIH), Capital Solution Design LLC (software used in a study funded by CAMH Foundation), and HAPPYneuron (software used in a study funded by Brain Canada). Within the past three years, he has also been an unpaid consultant to Myriad Neuroscience. N.H. Neufeld has received grant support from the Brain and Behavior Research Foundation, Canadian Institutes of Health Research, Physicians' Services Incorporated Foundation, and University of Toronto. L.D. Oliver receives grant support from the Brain and Behavior Research Foundation. J.D. Power has no disclosures. A.J. Rothschild has received grant or research support from Janssen, Otsuka, and the Irving S. and Betty Brudnick Endowed Chair in Psychiatry; is a consultant to Daiichi Sankyo, Inc., Sage Therapeutics, Xenon Pharmaceuticals, and Neumora Therapeutics; and has received royalties for the Rothschild Scale for Antidepressant Tachyphylaxis (RSAT)*, Clinical Manual for the Diagnosis and Treatment of Psychotic Depression, American Psychiatric Press, 2009; The Evidence-Based Guide to Antipsychotic Medications, American Psychiatric Press, 2010; The Evidence-Based Guide to Antidepressant Medications, American Psychiatric Press, 2012, and from UpToDate*. J.A. Sirey has no disclosures. A.N. Voineskos has received funding from the NIMH, Canadian Institutes of Health Research, Canada Foundation for Innovation, CAMH Foundation, and the University of Toronto. E.M. Whyte has received grant support from the NIMH and HRSA. G.S. Alexopoulos has received NIMH grants and has served in the speakers bureau of Takeda, Lundbeck, Otsuka, Alergan, Astra/Zeneca, Sunovion. A.J. Flint has received grant support from the U.S. National Institutes of Health, Patient-Centered Outcomes Research Institute, Canadian Institutes of Health Research, Brain Canada, Ontario Brain Institute, Alzheimer's Association, AGE-WELL, the Canadian Foundation for Healthcare Improvement, and the University of Toronto.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

- Alexopoulos, G. S., Meyers, B. S., Young, R. C., Kalayam, B., Kakuma, T., Gabrielle, M., ... Hull, J. (2000). Executive dysfunction and long-term outcomes of geriatric depression. *Archives of General Psychiatry*, *57*, 285–290.
- Barnes, T. R. (1989). A rating scale for drug-induced akathisia. *British Journal of Psychiatry*, *154*, 672–676.
- Blumberger, D. M., Mulsant, B. H., Emeremni, C., Houck, P., Andreescu, C., Mazumdar, S., ... Meyers, B. S. (2011). Impact of prior pharmacotherapy on remission of psychotic depression in a randomized controlled trial. *Journal of Psychiatric Research*, *45*, 896–901.
- Buckman, J. E. J., Underwood, A., Clarke, K., Saunders, R., Hollon, S. D., Fearon, P., & Pilling, S. (2018). Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. *Clinical Psychology Review*, *64*, 13–38.
- Burcusa, S. L., & Iacono, W. G. (2007). Risk for recurrence in depression. *Clinical Psychology Review*, *27*, 959–985.
- Chekroud, A. M., Bondar, J., Delgado, J., Doherty, G., Wasil, A., Fokkema, M., ... Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*, 154–170.
- Coryell, W., Leon, A., Winokur, G., Endicott, J., Keller, M., Akiskal, H., & Solomon, D. (1996). Importance of psychotic features to long-term course in major depressive disorder. *American Journal of Psychiatry*, *153*, 483–489.
- Delis, D., Kaplan, E., & Kramer, J. (2001). *Delis-Kaplan executive function scale*. San Antonio, TX: The Psychological Corporation.
- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2001). *Structured clinical interview for DSM-IV-TR axis I disorders - patient edition (SCID-I/P)*. New York: Biometrics Research Department.
- Flint, A. J., Bingham, K. S., Neufeld, N. H., Alexopoulos, G. S., Mulsant, B. H., Rothschild, A. J., ... Meyers, B. S.; STOP-PD II Study Group (2021) Association between psychomotor disturbance and treatment outcome in psychotic depression: A STOP-PD II report. *Psychological Medicine* *26*, 1–7.
- Flint, A. J., Meyers, B. S., Rothschild, A. J., Whyte, E. M., Alexopoulos, G. S., Rudorfer, M. V., ... Mulsant, B. H., STOP-PD II Study Group (2019) Effect of continuing olanzapine versus placebo on relapse among patients with psychotic depression in remission: The STOP-PD II randomized clinical trial. *Journal of the American Medical Association* *322*, 622–631.
- Flint, A. J., Meyers, B. S., Rothschild, A. J., Whyte, E. M., Mulsant, B. H., Rudorfer, M. V., & Marino, P., STOP-PD II Study Group (2013) Sustaining remission of psychotic depression: Rationale, design and methodology of STOP-PD II. *BMC Psychiatry* *13*, 38–49.
- Folstein, M., Folstein, S., & McHugh, P. (1975). "Mini mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.
- Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., ... Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, *48*, 851–855.
- Georgieva, R., Chekroud, A. M., & Krystal, J. H. (2017). Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: An individual patient-level data meta-analysis. *The Lancet. Psychiatry*, *4*, 230–237.
- Guy, W. (1976). Clinical global impressions. *ECDEU Assessment manual for psychopharmacology* (pp. 217–222). Washington, DC, US: Dept. of Health, Education, and Welfare.
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology Neurosurgery and Psychiatry*, *23*, 56–62.
- Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W. A., & Beekman, A. T. F. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatrica Scandinavica*, *122*, 184–191.
- Jääskeläinen, E., Juola, T., Korpela, H., Lehtiniemi, H., Nietola, M., Korkeila, J., & Miettunen, J. (2018). Epidemiology of psychotic depression - systematic review and meta-analysis. *Psychological Medicine*, *48*, 905–918.
- Jorm, A. F. (2004). The informant questionnaire on cognitive decline in the elderly (IQCODE): A review. *International Psychogeriatrics*, *16*, 275–293.
- Klein, N. S., Holtman, G. A., Bockting, C. L. H., Heymans, M. W., & Burger, H. (2018). Development and validation of a clinical prediction tool to estimate

- the individual risk of depressive relapse or recurrence in individuals with recurrent depression. *Journal of Psychiatric Research*, 104, 1–7.
- Miller, M. D., Paradis, C. F., Houck, P. R., Mazumdar, S., Stack, J. A., Rifai, A. H., ... Reynolds, C. F. 3rd. (1992). Rating chronic medical illness burden in geropsychiatric practice and research: Application of the cumulative illness rating scale. *Psychiatry Research* 41, 237–248.
- Nietola, M., Heiskala, A., Nordström, T., Miettunen, J., Korkeila, J., & Jääskeläinen, E. (2018). Clinical characteristics and outcomes of psychotic depression in the Northern Finland Birth Cohort 1966. *European Psychiatry*, 53, 23–30.
- Oquendo, M. A., Baca-Garcia, E., Kartachov, A., Khait, V., Campbell, C. E., & Richards, M. (2003). A computer algorithm for calculating the adequacy of antidepressant treatment in unipolar and bipolar depression. *Journal of Clinical Psychiatry*, 64, 825–833.
- Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological Reports*, 10, 799–812.
- Parker, G., Hadzi-Pavlovic, D., Brodaty, H., Boyce, P., Mitchell, P., Wilhelm, K., ... Eysers, K. (1993). Psychomotor disturbance in depression: Defining the constructs. *Journal of Affective Disorders*, 27, 255–265.
- Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J. F., ... Berlin, M. T. (2019). A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *Journal of Affective Disorders*, 243, 503–515.
- Randolph, C. (1998). *The repeatable battery for the assessment of neuropsychological status*. San Antonio, TX: The Psychological Corporation.
- Reynolds, C. F. III, Dew, M. A., Pollock, B. G., Mulsant, B. H., Frank, E., Miller, M. D., ... Kupfer, D. J. (2006). Maintenance treatment of major depression in old age. *New England Journal of Medicine* 354, 1130–1138.
- Simpson, G. M., & Angus, J. W. (1970). A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica Suppl* 212(S212), 11–19.
- Song, Y., Qian, L., Sui, J., Greiner, R., Li, X. M., Greenshaw, A. J., ... Cao, B. (2023). Prediction of depression onset risk among middle-aged and elderly adults using machine learning and Canadian Longitudinal Study on Aging cohort. *Journal of Affective Disorders*, 339, 52–57. doi: 10.1016/j.jad.2023.06.031
- Spitzer, R. L., & Endicott, J. (1979). *Schedule for affective disorders and schizophrenia* (3rd ed.). New York: New York State Psychiatric Institute, Biometrics Research.
- Stekhoven, D. J., & Buehlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, 28, 112–118.
- Williams, J. B. W., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., ... Kalali, A. (2008). The GRID-HAMD: Standardization of the Hamilton depression rating scale. *Int Clin Psychopharmacology*, 23, 120–129.
- Wojnarowski, C., Firth, N., Finegan, M., & Delgado, J. (2019). Predictors of depression relapse and recurrence after cognitive behavioural therapy: A systematic review and meta-analysis. *Behavioural Cognitive Psychotherapy*, 47, 514–529.
- Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67, 361–370.