JAMA Psychiatry | Original Investigation

Reevaluating the Efficacy and Predictability of Antidepressant Treatments
A Symptom Clustering Approach

Adam M. Chekroud, MSc; Ralitza Gueorguieva, PhD; Harlan M. Krumholz, MD, SM; Madhukar H. Trivedi, MD; John H. Krystal, MD; Gregory McCarthy, PhD

IMPORTANCE Depressive severity is typically measured according to total scores on questionnaires that include a diverse range of symptoms despite convincing evidence that depression is not a unitary construct. When evaluated according to aggregate measurements, treatment efficacy is generally modest and differences in efficacy between antidepressant therapies are small.

OBJECTIVES To determine the efficacy of antidepressant treatments on empirically defined groups of symptoms and examine the replicability of these groups.

DESIGN, SETTING, AND PARTICIPANTS Patient-reported data on patients with depression from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (n = 4039) were used to identify clusters of symptoms in a depressive symptom checklist. The findings were then replicated using the Combining Medications to Enhance Depression Outcomes (CO-MED) trial (n = 640). Mixed-effects regression analysis was then performed to determine whether observed symptom clusters have differential response trajectories using intent-to-treat data from both trials (n = 4706) along with 7 additional placebo and active-comparator phase 3 trials of duloxetine (n = 2515). Finally, outcomes for each cluster were estimated separately using machine-learning approaches. The study was conducted from October 28, 2014, to May 19, 2016.

MAIN OUTCOMES AND MEASURES Twelve items from the self-reported Quick Inventory of Depressive Symptomatology (QIDS-SR) scale and 14 items from the clinician-rated Hamilton Depression (HAM-D) rating scale. Higher scores on the measures indicate greater severity of the symptoms.

RESULTS Of the 4706 patients included in the first analysis, 1722 (36.6%) were male; mean (SD) age was 41.2 (13.3) years. Of the 2515 patients included in the second analysis, 855 (34.0%) were male; mean age was 42.65 (12.17) years. Three symptom clusters in the QIDS-SR scale were identified at baseline in STAR*D. This 3-cluster solution was replicated in CO-MED and was similar for the HAM-D scale. Antidepressants in general (8 of 9 treatments) were more effective for core emotional symptoms than for sleep or atypical symptoms. Differences in efficacy between drugs were often greater than the difference in efficacy between treatments and placebo. For example, high-dose duloxetine outperformed escitalopram in treating core emotional symptoms (effect size, 2.3 HAM-D points during 8 weeks, 95% CI, 1.6 to 3.1; $P < .001$), but escitalopram was not significantly different from placebo (effect size, 0.03 HAM-D points; 95% CI, −0.7 to 0.8; $P = .94$).

CONCLUSIONS AND RELEVANCE Two common checklists used to measure depressive severity can produce statistically reliable clusters of symptoms. These clusters differ in their responsiveness to treatment both within and across different antidepressant medications. Selecting the best drug for a given cluster may have a bigger benefit than that gained by use of an active compound vs a placebo.

JAMA Psychiatry. 2017;74(4):370-378. doi:10.1001/jamapsychiatry.2017.0025
Published online February 22, 2017.

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Adam M. Chekroud, MSc, Department of Psychology, Yale University, 2 Hillhouse Ave, New Haven, CT 06511 (adam.chekroud@yale.edu).

Meta-analyses[1] and factor analytic studies of large populations with depression[2,3] indicate that the symptoms of major depressive disorder are organized into 2 to 5 clusters depending on the checklist used. Nevertheless, clinical trials of patients with depression nearly always report total symptom severity scores as their primary outcome measures. These studies also frequently report the proportion of patients whose total symptom severity falls below a certain threshold and thus achieve clinical response or remission.[4] Few patients reach remission with their initial treatment, although depression eventually remits in most patients after a largely trial-and-error treatment selection process.[5] Statistical models might improve clinical outcomes by accelerating the treatment matching process. Despite concerted efforts using genomic data,[6] structural and functional magnetic resonance imaging,[7] and machine learning of clinical data,[8] performance in predicting outcomes remains modest.[9,10]

Heterogeneity among depressive symptoms may impede the evaluation of treatments for depression.[11,12] For example, treatment efficacy for one group of symptoms may be masked by a lack of efficacy for other symptoms, potentially explaining mixed results from large comparative efficacy meta-analyses.[4,13] For example, selective serotonin reuptake inhibitors are generally effective in reducing low mood[14] relative to other symptoms. However, evaluating outcomes on an individual symptom level may be cumbersome since clinicians would need to remember treatment guidelines specific to each symptom. Although symptoms might be grouped based on clinical experience (eg, "melancholic depression")[15] or the use of rating subscales (eg, Hamilton Rating Scale for Depression-7), novel associations might be overlooked by this process.

Statistical methods enable one to categorize depressive symptoms into subcomponents. For example, one study showed that nortriptyline hydrochloride is more effective than escitalopram in treating a neurovegetative symptom dimension, but escitalopram was more effective in treating mood and cognitive symptom dimensions.[16] However, traditional statistical approaches have some shortcomings. Factor analyses, for example, may generate complicated combinations of symptoms within particular dimensions.[16] These analyses also may be susceptible to experimenter bias since one often has to choose the desired number of clusters or components in the data, as in k means clustering.[17] By contrast, hierarchical clustering is an easy-to-visualize, deterministic method in which each symptom is assigned to a single cluster (ie, not loading across multiple clusters) without prespecifying the desired the number of clusters.

In this study, we explored the efficacy and predictability of antidepressant therapies in treating specific groups of symptoms (eMethods [which includes eTables 1-10 of various analyses] and eFigure 1 in the Supplement). We used an unsupervised machine-learning approach (hierarchical clustering) to establish a data-driven grouping of baseline symptoms. The clustering method was applied to patients from a large multisite trial of depression and a replication sample from an independent clinical trial with similar inclusion criteria. Next, we reanalyzed treatment outcomes for 9 archival clinical trials (**Table 1**) according to the severity of each symptom cluster (rather than total severity) to determine whether symptom clusters are equally responsive to antidepressant treatments and whether certain drugs and doses are more effective than others. Finally, we used supervised machine learning to predict outcomes specific to each cluster of symptoms since there may be good clinical or biological indicators of changes in some symptoms that do not correlate strongly with changes in other features of depression.

Key Points

Question Are antidepressants equally good at treating different kinds of symptoms in depression?

Findings Individual patient data from 9 clinical trials of major depression in 7221 patients were analyzed, with a focus on specific clusters of symptoms rather than total depressive severity. For each cluster, significant differences in efficacy between antidepressants were identified.

Meaning Antidepressant medications can be selected to benefit specific clusters of symptoms in depression.

Methods

Clinical Trial Data

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial is the largest prospective, randomized clinical trial of outpatients with major depressive disorder.[18-21] Eligible participants were treatment-seeking outpatients with a primary clinical (*DSM-IV*) diagnosis of nonpsychotic major depressive disorder scored 14 or higher on the 17-item Hamilton Depression (HAM-D) rating scale, were aged 18 to 75 years, and were recruited from primary and psychiatric care settings in the United States from June 2001 to April 2004.[19] We focused on the first treatment stage consisting of a 12-week course of citalopram hydrobromide. The present study was conducted from October 28, 2014, to May 19, 2016. It was approved by the Yale University Human Subjects Committee, with a waiver of informed consent.

The Combining Medications to Enhance Depression Outcomes (CO-MED) trial was a multisite, single-blind, randomized clinical trial comparing the efficacy of medication combinations in the treatment of unipolar major depressive disorder.[22,23] Eligible patients were aged 18 to 75 years, had a primary *DSM-IV*-based diagnosis of nonpsychotic major depressive disorder, had recurrent or chronic depression (current episode ≥2 years), scored 16 or higher on the 17-item HAM-D rating scale, and enrolled participants between March 2008 and February 2009. Patients were randomly allocated (1:1:1) to escitalopram plus placebo (monotherapy), escitalopram plus bupropion hydrochloride, or venlafaxine hydrochloride plus mirtazapine.

We also analyzed all arms from 7 randomized, multicenter, double-blind, placebo-controlled, and active comparator-controlled clinical trials of duloxetine for major depressive disorder (Table 1). Four different protocols were used for these

Table 1. Individual Patient-Level Data Aggregated From 9 Trials of Antidepressant Efficacy for Unipolar Major Depression

Protocol	Sample Size (N = 7221)	Treatment	Dose
STAR*D phase 1	4041	Citalopram	20-60 mg once daily
CO-MED	224	Escitalopram plus placebo	10-20 mg once daily
	221	Escitalopram plus bupropion extended release	Escitalopram, 10-20 mg once daily; bupropion extended release, 150-200 mg twice daily
	220	Venlafaxine extended release plus mirtazapine	Venlafaxine extended release, 37.5-300 mg once daily; mirtazapine, 15-45 mg once daily
HMAQ part A	70	Duloxetine	20-60 mg twice daily
	33	Fluoxetine	20 mg once daily
	70	Placebo	NA
HMAQ part B	82	Duloxetine	20-60 mg twice daily
	37	Fluoxetine	20 mg once daily
	74	Placebo	NA
HMAT part A	91	Duloxetine	20 mg twice daily
	84	Duloxetine	40 mg twice daily
	89	Paroxetine	20 mg once daily
	90	Placebo	NA
HMAT part B	86	Duloxetine	20 mg twice daily
	91	Duloxetine	40 mg twice daily
	87	Paroxetine	20 mg once daily
	89	Placebo	NA
HMAY part A	95	Duloxetine	40 mg twice daily
	93	Duloxetine	60 mg twice daily
	86	Paroxetine	20 mg once daily
	93	Placebo	NA
HMAY part B	93	Duloxetine	40 mg twice daily
	103	Duloxetine	60 mg twice daily
	97	Paroxetine	20 mg once daily
	99	Placebo	NA
HMCR	273	Duloxetine	60 mg twice daily
	273	Escitalopram	10 mg once daily
	137	Placebo	NA

Abbreviations: CO-MED, Combining Medications to Enhance Depression Outcomes; NA, not applicable; STAR*D, Sequenced Treatment Alternatives to Relieve Depression.

studies; parts A and B reflect trials run in parallel following the same protocol. All studies incorporated double-blind, variable-duration placebo lead-in periods. Safety and efficacy results from these studies have been published previously[24-27] and summarized as pooled analyses of safety[28] and efficacy.[29] Study HMCR is registered at clinicaltrials.gov.[30] The other studies were conducted before clinical trial registration was necessary.

Outcomes for STAR*D and CO-MED are based on the 16-item self-report Quick Inventory of Depressive Symptomatology (QIDS-SR) checklist during 12 weeks of treatment. Outcomes for all other trials are based on the 17-item HAM-D rating scale[31] during 8 weeks. We excluded the HAM-D "loss of insight" item because there is no equivalent in the QIDS-SR and excluded weight/appetite items because they were not collected in the same way across trials and are often excluded from item-level analyses[32] (eFigure 2 in the Supplement). Study selection was driven primarily by access to individual patient-level data. Patients provided informed consent to treatment when they participated in the original clinical trials. Consent was not needed for the present analyses since the data were deidentified. Of the 4706 patients included in the first analysis, 1722 (36.6%) were male; mean (SD) age was 41.2 (13.3) years. Of the 2515 patients included in the second analysis, 855 (34.0%) were male; mean age was 42.65 (12.17) years.

Symptom Clustering

Rating scales in depression include a diverse range of symptoms. We applied a data-driven approach to identify groups of symptoms within depression rating scales. Higher scores on the rating scales indicate more severe symptoms. Hierarchical clustering shows structure in data without making assumptions about the number of clusters that are present in the data and gives a deterministic solution. We applied agglomerative (bottom-up) hierarchical clustering to the QIDS-SR checklist completed at baseline in STAR*D by 4017 patients and replicated the analysis using baseline QIDS-SR data from CO-MED (n = 640) and the baseline HAM-D scale that was also collected on 4039 patients in STAR*D. We conducted multiple sensitivity analyses using alternative approaches (eFigures 3-9 and eTables 1-3 in the Supplement).

Evaluation of Treatment Outcomes

Treatment Efficacy

We analyzed the full intent-to-treat samples in all trials using linear mixed-effects regression models (STAR*D, 4041; CO-MED, 665; and other trials, 2515). The dependent measure was mean within-cluster severity: for each patient at each time point, we calculated the mean symptom severity within each cluster. Fixed effects included symptom cluster, time (log-transformed weeks), treatment regimen, and all 2- and 3-way interaction effects. We included a separate random intercept and slope for each symptom cluster with unstructured variance-covariance of the random effects within subject based on improvements in the Schwarz-Bayesian information criterion.[33] False-discovery rate-adjusted[34] *P* values were used to determine statistical significance for post hoc comparisons by cluster and drug within each mixed-model analysis.

One model was used to analyze QIDS-SR–based clusters across STAR*D and CO-MED, and another model was used to analyze HAM-D–based clusters for the 7 other placebo-controlled trials. In the HAM-D model, we also included the main effect of the trial to control for potential systematic differences among trials. Preliminary analyses of the 4 duloxetine doses in each cluster indicated that 120-mg/d and 80-mg/d dosages were not significantly different from each other but differed from the lower doses and placebo (eResults and eFigure 10 in the Supplement). The 60-mg/d and 40-mg/d duloxetine dosages were similar to each other and nearly indistinguishable from placebo. We therefore grouped cohorts into high-dose duloxetine (80-120 mg/d) and low-dose duloxetine (40-60 mg/d).

Outcome Predictability

We used a recently developed statistical modeling pipeline[8] to predict treatment outcomes specific to each symptom cluster using information available at baseline. We extracted 164 items, including demographics, medical and psychiatric histories, and specific symptom items that were used as predictor variables (eTable 10 in the Supplement). Penalized logistic regression (elastic net[35,36]) was then used to identify the 25 variables that best predicted each cluster separately. These variables were then used to train machine-learning algorithms (gradient boosting machines[37,38]), resulting in a separate model for each symptom cluster, with each using 25 predictor variables. Predictability was measured as the percentage of variance explained in final cluster scores (ie, R^2) using 5 repeats of 10-fold cross-validation. The statistical significance of each model was assessed using a permutation test (eMethods in the Supplement). We trained models on patients with complete baseline data for whom a severity score was recorded after 12 or more weeks of treatment (n = 1962) to ensure adequate treatment duration. To externally validate our predictive models, they were applied without modification to predict final cluster scores in CO-MED treatment completers. Here, statistical significance was measured by a *P* value calculated for Pearson correlations between predicted outcomes and observed outcomes in each treatment group of CO-MED. We did not have comparable predictor data in the duloxetine trials; thus, predictive analyses were conducted only for STAR*D and CO-MED. For significance, permutation-based tests used an α level of .01, mixed-effects regressions used a false-discovery rate correction and then an α level of .05, and Pearson correlations used an α level of .05.

Predictive and clustering analyses were implemented in *R*, version 3.2.3 (R Foundation). Efficacy analyses were conducted using SAS, version 9.4 (proc mixed) (SAS Institute).

Results

In 2 independent trials, we identified the same clustering of symptoms in the QIDS-SR checklist, consisting of core emotional, sleep (insomnia), and atypical symptoms (**Figure 1**A and B). A similar clustering solution was also found for the HAM-D scale checklist (Figure 1C). The clustering solution was robust across a number of sensitivity analyses using different parameters, time points, and approaches (eFigures 3-9 and eTables 1-3 in the Supplement).

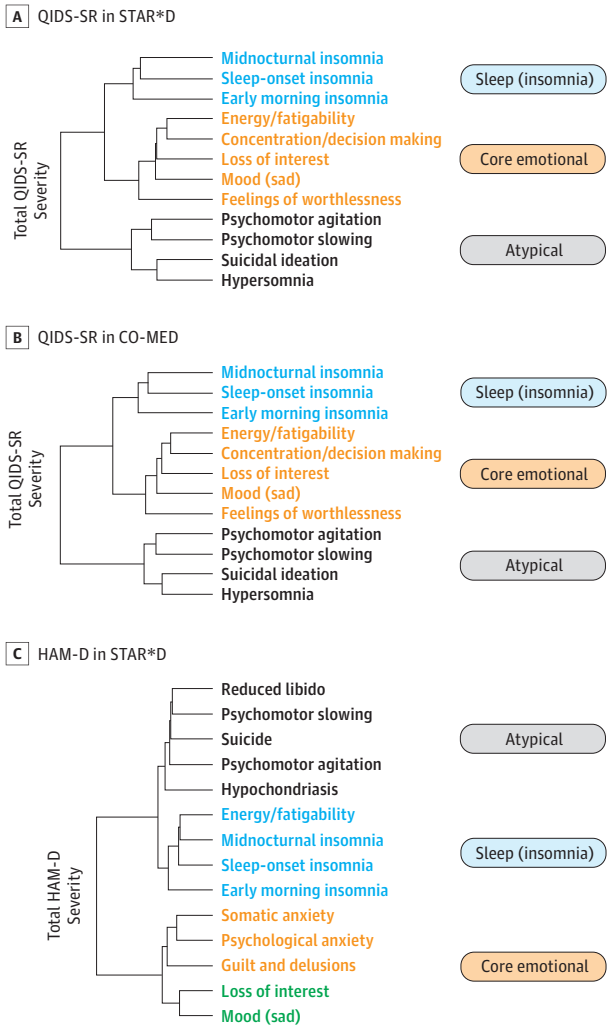
Efficacy Analyses

Treatment efficacy was measured according to the rate of symptom improvement over time (ie, steeper symptom trajectories are better, as shown in **Figure 2**). No antidepressant treatment worked equally well across all 3 symptom clusters. As shown in Figure 2A, when measured according to the QIDS-SR, trajectories were significantly better for core emotional symptoms than for either sleep symptoms or atypical symptoms for citalopram, escitalopram with placebo, and escitalopram with bupropion (all β>0.079; all false-discovery rate corrected $P < .001$). Sleep trajectories were also better than atypical trajectories for these 3 treatments (all β>0.099; all $P \leq .001$). As shown in Figure 2B, when measured according to the HAM-D rating scale, a similar pattern was observed. Core emotional trajectories were better than sleep and atypical trajectories for all treatments (all β>0.12; all $P \leq .001$). Sleep trajectories were also better than atypical trajectories for low-dose duloxetine and escitalopram (all β>0.080; all $P \leq .001$). All slope contrast estimates, SEs, 95% CIs, and *P* values are included in eTables 4 and 5 in the Supplement.

To interpret the magnitude of differences between drugs, we calculated an effect size (ES), measured in raw rating scale points, that reflects the difference between treatments in reducing the overall severity of a symptom cluster (ie, we multiplied slope contrasts by the natural log of treatment duration and then by the number of symptoms in each cluster). For example, in this study, high-dose duloxetine was significantly better than escitalopram in treating atypical symptoms, such that a patient's total improvement in atypical severity was a mean of 1.9 HAM-D points greater with high-dose duloxetine than escitalopram (ES, 1.9; 95% CI, 1.4-2.3; false-discovery rate corrected $P < .001$).

For each symptom cluster, there were significant differences in efficacy between treatments (Figure 2). Combined escitalopram and bupropion treatment was significantly more effective in treating core emotional symptoms than citalopram (ES, 0.7 QIDS-SR points; 95% CI, 0.2 to 1.3; $P = .03$). For sleep/insomnia symptoms, venlafaxine with mirtazapine outperformed citalopram (ES, 1.4; 95% CI, 1.0 to 1.8; $P < .001$). For

Figure 1. Data-Driven Decomposition of Depressive Checklists Using Hierarchical Clustering



This procedure sequentially groups symptoms according to the similarity of their responses across a patient cohort. With this procedure, groups of symptoms that merge at high values relative to the merge points of their subgroups are considered candidates for natural clusters. A and B, In the Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR) checklist, we identified an identical 3-cluster solution in both the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) (n = 4017) and Combining Medications to Enhance Depression Outcomes (CO-MED) trials (n = 640). C, A comparable symptom structure was also observed at baseline for STAR*D patients when measured according to the Hamilton Depression (HAM-D) rating scale. The names of the individual checklist items are colored according to their cluster assignment. Line lengths in the dendogram reflect how similar items or clusters are to one another (shorter line length indicates greater similarity).

core emotional symptoms in HAM-D scale trials (Figure 2B), high-dose duloxetine outperformed escitalopram (ES, 2.3 HAM-D points; 95% CI, 1.6 to 3.1; $P < .001$). Escitalopram was not significantly different from placebo for core emotional symptoms (ES, 0.03 HAM-D points; 95% CI, −0.7 to 0.8; $P = .94$). For sleep symptoms, high-dose duloxetine outperformed fluoxetine (ES, 0.9; 95% CI, 0.1 to 1.7; $P = .046$). For atypical symptoms, high-dose duloxetine outperformed all others (ES, 0.5-1.9) and escitalopram was worse than placebo (ES, 0.7; 95% CI, 0.3 to 1.1; $P = .002$). Among our HAM-D studies, only 2 antidepressant treatments (high-dose duloxetine and paroxetine) outperformed placebo for all 3 symptom clusters. All other comparisons are presented in eTables 6 and 7 in the Supplement.

Predictive Analyses

Within STAR*D, although all models performed significantly above chance (all $P < .01$), we observed substantial variability in the predictability of outcomes for each cluster (**Table 2** and eTable 8 in the Supplement). The sleep symptom cluster was the most predictable ($R^2 = 19.6\%$; SD, 5.0%; $P < .01$) and substantially more predictable than core symptoms ($R^2 = 14.5\%$; SD, 4.6%; $P < .01$) and atypical symptoms ($R^2 = 15.1\%$; SD, 5.3%; $P < .01$). The observed range in cluster predictability (R^2 difference, 5.1%) was also significantly larger than any range observed during permutation testing (mean [SD] range, 0.56% [0.50%]; $P < .01$). We inspected the best predictive baseline variables for each model separately, highlighting those identified as predictive for 1 cluster but not others (ie, specific predictors) (Table 2). Baseline HAM-D scale severity was a top predictor of core emotional outcomes but not any of the other 3 clusters. Baseline atypical symptom severity and hypersomnia predicted atypical outcomes; baseline sleep cluster severity and early-morning insomnia predicted sleep outcomes.

We then applied the best-performing models, without modification, to predict outcomes for each cluster in the 3 treatment groups of CO-MED (**Figure 3**). Performance was statistically above chance, although clinically modest, for predicting core emotional outcomes in the escitalopram monotherapy arm ($r_{149} = 0.18$; $P = .03$) and the venlafaxine-mirtazapine arm ($r_{138} = 0.17$; $P = .04$). Performance was above chance predicting sleep outcomes in the escitalopram-bupropion arm ($r_{132} = 0.36$; $P < .001$).

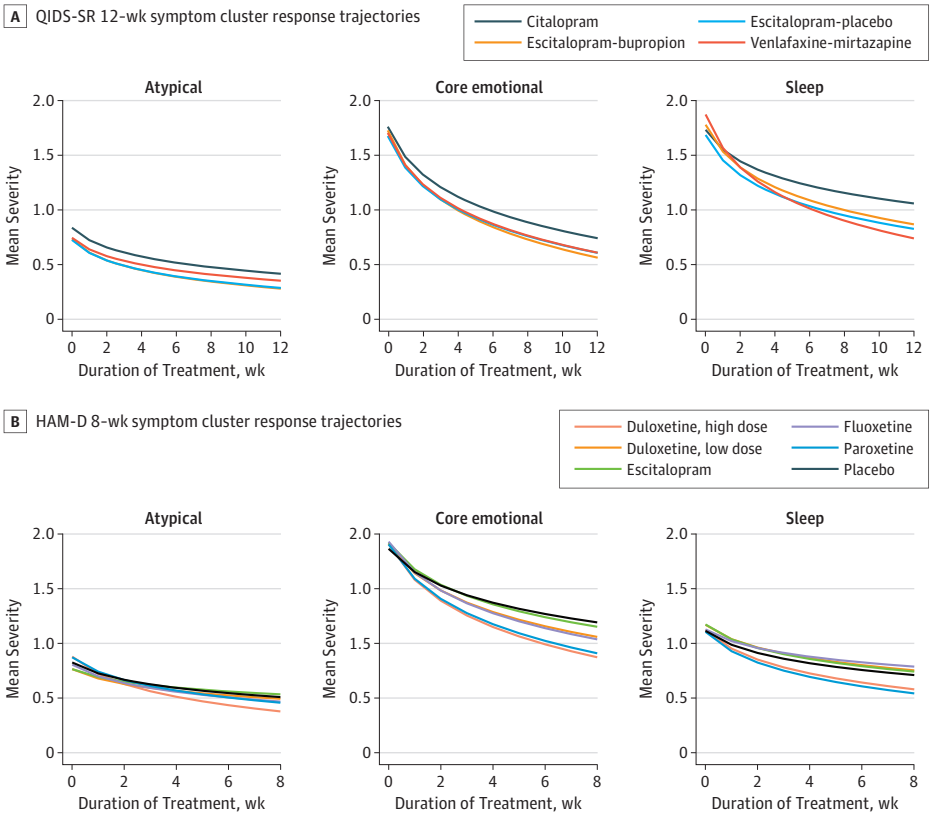
Clinical Decision Support Tool

To help translate these findings into clinical practice, we based a clinical decision support tool on these findings. It is implemented as a brief questionnaire that can be accessed from any web browser and returns results in real time (https://www.spring.care/spring-assessment).

Discussion

Using a data-driven approach, we identified 3 symptom clusters within the QIDS-SR checklist. We replicated our clustering solution in an independent trial cohort (CO-MED) and found it to be robust across different parameters and time points and consistent with other statistical approaches. No antidepressant was equally effective for all 3 symptom clusters, and, for each symptom cluster, there were significant differences in treatment efficacy between drugs. Antidepressants in general worked best in treating core emotional and sleep symptoms and were less effective in treating atypical symptoms. The magnitude of these differences suggests that selecting the best drug for a given cluster may have a bigger benefit than that

Figure 2. Model-Fitted Outcome Trajectories for Each Symptom Cluster



A, Measured according to the Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR) checklist in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) and Combining Medications to Enhance Depression Outcomes (CO-MED) trials (12 weeks). B, Measured according to the Hamilton Depression (HAM-D) rating scale in 7 phase 3, placebo-controlled trials of duloxetine (8 weeks). The y-axes represent mean severity within a cluster and so should be multiplied by the number of symptoms within a cluster to convert to original units.

gained by use of an active compound vs a placebo. Treatment outcomes at the symptom cluster level were predictable by machine learning of self-report data.

These results might help to guide future research on personalized antidepressant treatment. The 2015 revision to the 2008 British Association for Psychopharmacology guidelines indicate that clinicians should "match choice of antidepressant to individual patient requirements…taking into account likely short-term and long-term effects."[39(p463)] However, there is currently little appropriately powered evidence on which symptom-specific recommendations might be made. Our present finding of better trajectories for core symptoms with citalopram supports Genome-Based Therapeutic Drugs for Depression findings that mood and cognitive symptom dimensions were significantly better for escitalopram treatment than nortryptyline.[16] Whereas large-scale comparative efficacy studies of aggregate severity show modest (if any) differences between antidepressants,[4,13] our results at the symptom cluster level indicate substantial differences between drugs both within and across putative antidepressant classes. Moving forward, we must establish how improvements in a given cluster relate to quality of life, keeping in mind that medication tolerability remains an important clinical concern (as reviewed elsewhere[4,13]).

The approach outlined in this article may have implications for the drug approval process. United States Food and Drug Administration and European Medicines Association approval are currently determined in trials that use aggregate scores on severity measures to enroll patients or measure outcomes. Although some trials have used a specific symptom as an outcome (eg, depressed mood), our findings indicate that medications might be developed for specific clusters of symptoms, as they appear to respond differentially to antidepressant medications. Symptom clusters may also enable drug testing in smaller but more informative populations with a more consistent phenotype. This approach is consistent with the National Institute of Mental Health Research Domain Criteria[40]—symptom clusters or dimensions might have distinctive underlying neural circuitry and signaling mechanisms—and paves the way for developing treatments that target and biomarkers that predict changes in specific clusters of symptoms.

Further clinical research will determine whether these clusters generalize to other cohorts and reflect good candidates for a true symptom structure in major depression.[41,42] The present cluster structure resembles that of other scales in other large samples of patients with depression,[1-4] although a recent review concluded that the debate is not over.[41] These studies and ours are largely consistent in isolating symptoms of insomnia, a core group of symptoms that includes low mood, anhedonia, and low self-worth. However, direct comparisons are impeded by the use of many different rating scales in depression.[42] Our data-driven approach offers some novel symptom groupings relative to previous approaches. For instance, our emotional cluster resembled the HRSD-7

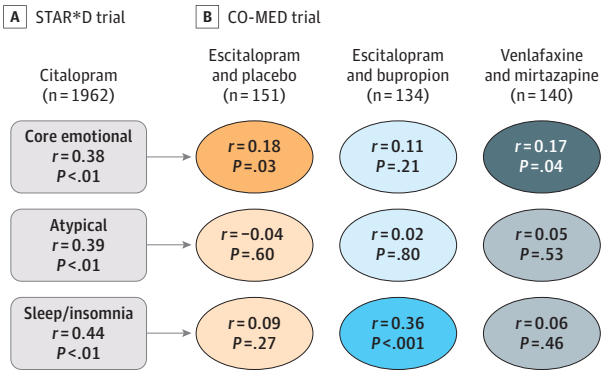
Table 2. Variables That Are Most Predictive of Cluster-Level Outcomes in Major Depressive Disorder

Core Emotional Cluster	Atypical Cluster	Sleep/Insomnia Cluster
Baseline core symptoms	Baseline atypical symptoms[a]	Baseline sleep symptoms[a]
Employed?	QIDS hypersomnia[a]	Age
Initial HAM-D severity[a]	Employed?	Black/African American
QIDS psychomotor agitation	QIDS psychomotor agitation	Employed?
QIDS energy/fatigability	Anxious standing in long lines	QIDS psychomotor agitation
Family thinks patient has drug problem	Sex[a]	QIDS sleep-onset insomnia
Years of education	Baseline core symptoms	QIDS early-morning insomnia[a]
Black/African American	Marriage problems because of drinking[a]	Frequent worrying about writing in public
Anxious about driving/riding in car	Thought might go crazy/lose control during anxiety attack[a]	Bothered by bodily aches and pains
Ever witnessed a traumatic event	Bothered by bodily aches and pains	Anxious about being in crowded places
HAM-D loss of insight	QIDS suicidal ideation	No. of previous depressive episodes[a]

Abbreviations: HAM-D, Hamilton Depression rating scale; QIDS, Quick Inventory of Depressive Symptomatology.

[a] Predictors that were not among the top 25 predictors for any other cluster (ie, cluster-specific predictors).

Figure 3. Use of Machine Learning to Predict Outcomes Specific to Each Symptom Cluster



For each symptom cluster, a new model was trained on patients who received citalopram in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (A). After cross-validation, we applied the models to patients in 3 treatment arms of the Combining Medications to Enhance Depression Outcomes (CO-MED) trial (B) to test their ability to generalize to an independent clinical trial sample. Core emotional symptoms could be predicted with significantly above-chance performance in the escitalopram with placebo and venlafaxine with mirtazapine arms. Sleep/insomnia symptoms could be predicted above chance for escitalopram with bupropion.

subscale but never included a suicide item, and when scored according to the HAM-D scale, the HRSD-7 energy/fatigability item clustered with insomnia symptoms rather than emotional symptoms. There were slight differences between the QIDS-SR and HAM-D scale results. In the HAM-D scale, the emotional cluster included an anxiety item, whereas in the QIDS-SR scale, the same cluster included low energy and concentration. The energy/concentration item falls in the sleep cluster for the HAM-D scale. This data-driven approach may have identified a set of symptoms in the emotional presentation of depression that may have neural circuit correlates that are more cohesive than either the *DSM* criteria or theory-driven clusters, such as the Bech/Maier scales, which have not yet produced meaningful signatures on neural circuits or treatment response prediction.[10,43] Finally, the atypical cluster contains items that are not considered atypical items in the *DSM*, so conclusions about broader atypical symptoms should not be drawn from the naming of this cluster.

Limitations

This study has some limitations. First, there was a high degree of study heterogeneity. Two rating scales (clinician-rated HAM-D vs self-rated QIDS-SR) and treatment durations (8 vs 12 weeks) were used. The studies used a mixture of fixed- and variable-dosage protocols and had differences in blinding (STAR*D was unblinded, CO-MED was single-blind, and all other trials were double-blind). The consistency of these findings from 7000 patients from these heterogeneous studies suggests that the findings should generalize. However, study differences precluded direct comparisons using all available data, and study selection based on data availability may be a source of bias.[44] Our inclusion of placebo-controlled duloxetine trials was critical for considering the pattern of cluster response trajectories for placebo and determining whether trajectories were better with drug treatment than placebo. Ideally, behavioral interventions might be focused on atypical symptoms that are generally less responsive to antidepressants or combined with other focused interventions for specific/residual symptoms (eg, modafinil for energy/fatigue, zolpidem for insomnia). Finally, group-level differences do not translate to individual patient differences in a simple manner[45]; therefore, further research is needed to test whether the web tool is accurate and effective in real-world practice.

Larger limitations surround the interpretation of current predictive analyses (eTable 9, eFigure 11, and eDiscussion in the Supplement). Generalizability of our original pipeline was poor. Alternative analytic strategies may be more effective (eFigure 10 in the Supplement). This limitation highlights the importance of externally validating predictive tools rather than relying on metrics based on the discovery sample.[8] Because it is impractical for each model to require 25 different items, we must identify a more limited group of predictor variables to use cluster-specific tools in the clinic.

Conclusions

Clusters of symptoms are detectable in 2 common depression rating scales, and these symptom clusters vary in their responsiveness to different antidepressant treatments. These patterns may offer clinicians evidence for tailoring antidepressant selection according to the symptoms that a specific patient is experiencing immediately—almost doubling the expected effect size of a treatment.

ARTICLE INFORMATION

Accepted for Publication: December 31, 2016.

Published Online: February 22, 2017. doi:10.1001/jamapsychiatry.2017.0025

Author Affiliations: Department of Psychology, Yale University, New Haven, Connecticut (Chekroud, McCarthy); Spring Health, New York City, New York (Chekroud); Center for Outcomes Research and Evaluation, Yale–New Haven Hospital, New Haven, Connecticut (Chekroud, Krumholz); Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut (Gueorguieva); Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut (Krumholz); Department of Health Policy and Management, Yale School of Public Health, New Haven, Connecticut (Krumholz); Department of Psychiatry, University of Texas–Southwestern Medical School, Dallas (Trivedi); Department of Psychiatry, Yale University, New Haven, Connecticut (Krystal).

Author Contributions: Drs Krystal and McCarthy contributed equally to the study. Mr Chekroud and Dr Gueorguieva had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Chekroud, Krystal.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Chekroud, Gueorguieva.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Chekroud, Gueorguieva.
Obtained funding: Chekroud, Krystal.
Administrative, technical, or material support: Chekroud, Krystal.
Study supervision: Chekroud, Krystal, McCarthy.

Conflict of Interest Disclosures: Mr Chekroud holds equity in Spring Health (doing business as Spring Care Inc), a behavioral health startup. He is lead inventor on a provisional patent submission by Yale University. Dr Gueorguieva discloses consulting fees for Palo Alto Health Sciences and Mathematica Policy Research and a provisional patent submission by Yale University (Y0087.70116US00). Dr Krumholz is a recipient of research agreements from Medtronic and Janssen (a pharmaceutical company of Johnson & Johnson), through Yale University, to develop methods of clinical trial data sharing; is the recipient of a grant from the US Food and Drug Administration and Medtronic to develop methods for postmarket surveillance of medical devices; works under contract with the Centers for Medicare & Medicaid Services to develop and maintain performance measures; chairs (paid) a cardiac scientific advisory board for UnitedHealth; and is the founder of Hugo, a personal health information platform. Dr Trivedi has served as a paid adviser or consultant to Abbott, Abdi Ibrahim, Akzo (Organon), Alkermes, AstraZeneca, Axon Advisors, Bristol-Myers Squibb, Cephalon, Cerecor, Concert Pharmaceuticals, Eli Lilly, Evotec, Fabre Kramer Pharmaceuticals, Forest Pharmaceuticals, GlaxoSmithKline, Janssen Global Services, Janssen Pharmaceutical Products, Johnson & Johnson PRD, Libby, Lundbeck, Mead Johnson, MedAvante, Medtronic, Merck, Mitsubishi Tanabe Pharma Development America, Naurex, Neuronetics, Otsuka, Pamlab, Parke-Davis, Pfizer, PgxHealth, Phoenix Marketing Solutions, Rexahn Pharmaceuticals, Ridge Diagnostics, Roche Products, Sepracor, Shire Development, Sierra, SK Life and Science, Sunovion, Takeda, Tal Medical/Puretech Venture, Targacept, Transcept, VantagePoint, Vivus, and Wyeth-Ayerst Laboratories; he has received research support from the Agency for Healthcare Research and Quality, Corcept Therapeutics, Cyberonics, National Alliance for Research on Schizophrenia and Depression (now The Brain & Behavior Research Foundation), National Institute of Mental Health (NIMH), National Institute for Drug Abuse, Novartis, Pharmacia & Upjohn, Predix Pharmaceuticals (Epix), and Solvay. Dr Krystal is the editor of *Biological Psychiatry*. He has been a paid consultant to the following companies: LLC, AstraZeneca Pharmaceuticals, Biogen, Biomedisyn Corporation, Forum Pharmaceuticals, Janssen Pharmaceuticals, Orsuka America Pharmaceutical, Sunovion Pharmaceuticals, Takeda Industries, and Taisho Pharmaceutical Co. He is an unpaid member of the Scientific Advisory Board of Biohaven Pharmaceuticals, Blackthorn Therapeutics, Lohocla Research Corporation, Luc Therapeutices, Pfizer Pharmaceuticals, Spring Care, Inc, and TRImaran Pharma. He holds stock in ArRETT Neuroscience and Biohaven Pharmaceuticals and stock options in Blackthorn Therapeutics and Luc Therapeutics. Dr Krystal has the following patents and inventions: (1) dopamine and noradrenergic reuptake inhibitors in treatment of schizophrenia (patent No. 5,447,948); (2) co-inventor on a filed patent application by Yale University related to targeting the glutamatergic system for the treatment of neuropsychiatric disorders (PCTWO06108055A1); (3) intranasal administration of ketamine to treat depression (US application No. 14/197,767 and US application or Patient Cooperation Treaty international application No. 14/306,382); (4) composition and methods to treat addiction (provisional use patent application No. 61/973/961); and (5) treatment selection for major depressive disorder (US Patent and Trademark Office docket No. Y0087.70116US00). No other disclosures were reported.

Funding/Support: This study was supported in part by Yale University; The William K. Warren Foundation; grants 1UH2TR000960-01 and 5ULTR000142-08 from the National Center for Advancing Translational Science; the Department of Veterans Affairs (National Center for Posttraumatic Stress Disorder); grants P50AA12870 and M01RR00125 from the National Institute on Alcohol Abuse and Alcoholism; and grant UL1 RR024139 from the Yale Center for Clinical Investigation.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Data for Sequenced Treatment Alternatives to Relieve Depression (STAR*D) and Combining Medications to Enhance Depression Outcomes (CO-MED) were acquired from the NIMH through limited access data use certificates (eAppendix in the Supplement). Data for other trials were provided by Eli Lilly and Company. Amanda Zheutlin, MS (Yale University), Nikolaos Koutsouleris, MD (Ludwig-Maximilians University), and Martin Paulus, MD (Laureate Institute for Brain Research), provided advice and thoughtful comments on this article; there was no financial compensation.

REFERENCES

1. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J Clin Psychol*. 2006;62(1):123-146.

2. Li Y, Aggen S, Shi S, et al. The structure of the symptoms of major depression: exploratory and confirmatory factor analysis in depressed Han Chinese women. *Psychol Med*. 2013;44(7):1391-1401

3. Romera I, Delgado-Cohen H, Perez T, Caballero L, Gilaberte I. Factor analysis of the Zung self-rating depression scale in a large sample of patients with major depressive disorder in primary care. *BMC Psychiatry*. 2008;8:4.

4. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*. 2009;373(9665):746-758.

5. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am J Psychiatry*. 2006;163(11):1905-1917.

6. Schatzberg AF, DeBattista C, Lazzeroni LC, Etkin A, Murphy GM Jr, Williams LM. ABCB1 genetic effects on antidepressant outcomes: a report from the iSPOT-D trial. *Am J Psychiatry*. 2015;172(8):751-759.

7. Schmaal L, Marquand AF, Rhebergen D, et al. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. *Biol Psychiatry*. 2015;78(4):278-286.

8. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250.

9. Trivedi MH. Modeling predictors, moderators and mediators of treatment outcome and resistance in depression. *Biol Psychiatry*. 2013;74(1):2-4.

10. Trivedi MH, McGrath PJ, Fava M, et al. Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *J Psychiatr Res*. 2016;78:11-23.

11. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*. 2015;172:96-102.

12. Olbert CM, Gala GJ, Tupler LA. Quantifying heterogeneity attributable to polythetic diagnostic criteria: theoretical framework and empirical application. *J Abnorm Psychol*. 2014;123(2):452-462.

13. Gartlehner G, Hansen RA, Morgan LC, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. *Ann Intern Med*. 2011;155(11):772-785.

14. Hieronymus F, Emilsson JF, Nilsson S, Eriksson E. Consistent superiority of selective serotonin

reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. *Mol Psychiatry*. 2016;21(4):523-530.

15. Lin SY, Stevens MB. The symptom cluster-based approach to individualize patient-centered treatment for major depression. *J Am Board Fam Med*. 2014;27(1):151-159.

16. Uher R, Maier W, Hauser J, et al. Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. *Br J Psychiatry*. 2009;194(3):252-259.

17. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc C*. 1979; 28(1):100-108.

18. Rush AJ, Wisniewski SR, Warden D, et al. Selecting among second-step antidepressant medication monotherapies: predictive value of clinical, demographic, or first-step treatment features. *Arch Gen Psychiatry*. 2008;65(8):870-880.

19. Trivedi MH, Rush AJ, Wisniewski SR, et al; STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28-40.

20. Warden D, Rush AJ, Trivedi MH, Fava M, Wisniewski SR. The STAR*D Project results: a comprehensive review of findings. *Curr Psychiatry Rep*. 2007;9(6):449-459.

21. clinicaltrials.gov. Sequenced Treatment Alternatives to Relieve Depression (STAR*D). NCT00021528. https://clinicaltrials.gov/ct2/show/NCT00021528. Accessed January 13, 2017.

22. Rush AJ, Trivedi MH, Stewart JW, et al. Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *Am J Psychiatry*. 2011;168(7):689-701.

23. clinicalTrials.gov. Combining Medications to Enhance Depression Outcomes (CO-MED). NCT00590863. https://clinicaltrials.gov/ct2/show/NCT00590863. Accessed January 13, 2017.

24. Goldstein DJ, Mallinckrodt C, Lu Y, Demitrack MA. Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial. *J Clin Psychiatry*. 2002;63(3):225-231.

25. Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA. Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. *J Clin Psychopharmacol*. 2004;24(4):389-399.

26. Detke MJ, Wiltse CG, Mallinckrodt CH, McNamara RK, Demitrack MA, Bitter I. Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur Neuropsychopharmacol*. 2004;14(6):457-470.

27. Perahia DGS, Wang F, Mallinckrodt CH, Walker DJ, Detke MJ. Duloxetine in the treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur Psychiatry*. 2006;21(6):367-378.

28. Hudson JI, Wohlreich MM, Kajdasz DK, Mallinckrodt CH, Watkin JG, Martynov OV. Safety and tolerability of duloxetine in the treatment of major depressive disorder: analysis of pooled data from eight placebo-controlled clinical trials. *Hum Psychopharmacol*. 2005;20(5):327-341.

29. Mallinckrodt CH, Prakash A, Houston JP, Swindle R, Detke MJ, Fava M. Differential antidepressant symptom efficacy: placebo-controlled comparisons of duloxetine and SSRIs (fluoxetine, paroxetine, escitalopram). *Neuropsychobiology*. 2007;56(2-3):73-85.

30. clinicaltrials.gov. Duloxetine Compared to Escitalopram and Placebo in the Treatment of Patients With Depression. NCT00073411. https://clinicaltrials.gov/ct2/show/NCT00073411. Accessed January 13, 2017.

31. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.

32. van Borkulo C, Boschloo L, Borsboom D, Penninx BWJH, Waldorp LJ, Schoevers RA. Association of symptom network structure with the course of depression. *JAMA Psychiatry*. 2015;72(12):1219-1226.

33. Pauler D. The Schwarz criterion and related methods for normal linear models. *Biometrika*. 1998;85(1):13-27.[REMOVED IF= FIELD]

34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289-300.

35. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B*. 2005;67(2):301-320.

36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.

37. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 1999;38(3):367-378.

38. Ridgeway G. Generalized boosted models: a guide to the gbm package. *Compute (Greensboro)*. 2007;1(4):1-12.

39. Cleare A, Pariante CM, Young AH, et al; Members of the Consensus Meeting. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 2008 British Association for Psychopharmacology guidelines. *J Psychopharmacol*. 2015;29(5):459-525.

40. Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry*. 2010;167(7):748-751.

41. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time…or not? lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychol Assess*. 2016;28(11):1354-1367.

42. Fried EI. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J Affect Disord*. 2017;208:191-197.

43. Williams LM. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry*. 2016;3(5):472-480.

44. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221.

45. Snippe E, Simons CJP, Hartmann JA, et al. Change in daily life behaviors and depression: within-person and between-person associations. *Health Psychol*. 2016;35(5):433-441.