The Journal of Clinical Psychiatry

Supplementary Material

Article Title: Machine Learning Prediction of Quality of Life Improvement During Antidepressant Treatment of Patients With Major Depressive Disorder: A STAR*D and CAN-BIND-1 Report

Author(s): Tejas Phaterpekar, MDS; John-Jose Nunez, MD, MSc; Emma Morton, PhD; Yang S. Liu, PhD; Bo Cao, PhD; Benicio N. Frey, MD, PhD; Roumen V. Milev, MD, PhD; Daniel J. Müller, MD, PhD; Susan Rotzinger, PhD; Claudio N. Soares, MD, PhD; Valerie H. Taylor, MD, PhD; Rudolf Uher, MD, PhD; Sidney H. Kennedy, MD; and Raymond W. Lam, MD

DOI Number: https://doi.org/10.4088/JCP.23m14864

LIST OF SUPPLEMENTARY MATERIAL FOR THE ARTICLE

The Journal of Clinical Psychiatry

13. **Table 13** P-values for Random Forest balanced accuracy, comparing STAR*D internal validation vs CAN-BIND-1 external validation performance, across feature sets

14. **Appendix 1** Description of the overlapping features used in this work

15. **Appendix 2** Participant Selection

16. **Figure 1** A flowchart summary of model training and evaluation, where N represents sample size for each of the steps

17. **Figure 2** SHAP values taken from a randomly sampled Random Forest model, trained on the overlapping feature set (k = 100)

18. **Figure 3** SHAP values taken from a randomly sampled Random Forest model, trained on the full dataset (k= 480)

DISCLAIMER

This Supplementary Material has been provided by the author(s) as an enhancement to the published article. It has been approved by peer review; however, it has undergone neither editing nor formatting by in-house editorial staff. The material is presented in the manner supplied by the author.

Supplementary Table 1. Inclusion/exclusion criteria from the STAR*D and CAN-BIND-1 datasets.

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Cohort	Canadian Biomarker Integration Network in Depression-1 (CAN-BIND-1) Cohort
Inclusion Criteria	
• 18-75 years of age	• 18-60 years of age
• Outpatients	• Outpatients
• HRSD17 score >=14	• MADRS score >= 24
• DSM-IV criteria for single or recurrent nonpsychotic MDD	• DSM-IV-TR criteria for MDE in MDD
• Not currently taking citalopram or have been taking for < 7 days	• No psychotropic medications for at least 5 half-lives before baseline
Exclusion Criteria	
• Pregnant individuals	• Pregnant or breastfeeding individuals
• Psychosis in current or prior MDD episodes	• Psychosis in current episode
• Diagnosis of bipolar disorder	• Diagnosis of bipolar I or II disorder
• History of schizophrenia, schizoaffective disorder, psychosis, anorexia, bulimia, or obsessive compulsive disorder	• Has a different psychiatric diagnosis as the primary diagnosis
• Individuals who required immediate hospitalization for substance detoxification or treatment	• Individuals who have substance abuse/dependence in the past 6 months
• Individuals who have additional medical conditions or use medication that contraindicate any level 1 or 2 treatments.	• Individuals who have uncontrolled medical conditions, or significant neurological disorders/head trauma
• Individuals currently requiring mood stabilizers or antipsychotic drugs	• Individuals with a history of antidepressant-induced hypomania or showing any other risk factors for hypomanic switch on antidepressants
• Individuals who have experienced lack of response or clear intolerability to an adequate trial of an SSRI in their current MDD episode	• Individuals who have failed or had intolerance to a trial of escitalopram or aripiprazole, OR have failed 4 or more pharmacologic interventions.
• Individuals who did not respond to 16 or more cognitive therapy sessions, or 7 or more electroconvulsive therapy sessions, during current episode	• Individuals who have started psychological treatment in the past 3 months leading up to baseline, who intend to continue this modality
	• Individuals with contraindications to magnetic resonance imaging
	• Individuals with high suicidal risk or any significant personality disorder which might interfere with the treatment protocol (decided by clinical judgement)

Abbreviations: HRSD, Hamilton Rating Scale for Depression; MADRS, Montgomery Asberg Depression Rating Scale; MDE, major depressive episode; MDD, major depressive disorder; SSRI, selective serotonin reuptake inhibitor.

Supplementary Table 2. A comparison of feature types included in the various feature sets studied.

Feature Set Name::	Overlapping	Q-LES-Q-SF only	QIDS only	Q-LES-Q-SF + QIDS	No QIDS+ Q-LES-Q-SF
Features:	Demo + WPAI + Psychiatric Hx + Q-LES-Q + QIDS	Q-LES-Q	QIDS	QLES-Q-SF + QIDS	Demo + WPAI + Psychiatric Hx

Abbreviations: Hx, history; Demo: demographics; WPAI, Work Productivity and Impairment Q-LES-Q; Quality of Life Enjoyment and Satisfaction Questionnaire Short Form; QIDS, 16-item Quick Inventory of Depressive Symptomatology.

Supplementary Table 3. STAR*D internal validation (k = 100) results with additional metrics[a].

	Mean Scores for Test and Train								
Model	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	F1 Score
Dummy Classification	4.95 (19.51)	86.06 (338.93)	19.94 (81.07)	20.05 (80.49)	0.2 (0.2)	0.81 (0.81)	0.2 (0.19)	0.81 (0.81)	0.2 (0.19)
Random Forest	14.45 (78.27)	90.24 (350.53)	15.76 (69.47)	10.55 (21.73)	0.58 (0.78)	0.85 (0.83)	0.48 (0.53)	0.9 (0.94)	0.52 (0.63)
Logistic Regression	19 (84)	85 (330)	21 (90)	6 (16)	0.76 (0.84)	0.8 (0.79)	0.48 (0.48)	0.93 (0.95)	0.58 (0.61)
Elastic Net	19.28 (84.81)	85 (328.12)	21 (91.88)	5.72 (15.19)	0.77 (0.85)	0.8 (0.78)	0.48 (0.48)	0.94 (0.96)	0.59(0.61)
KNN	11 (100)	90 (420)	16 (0)	14 (0)	0.44 (1)	0.85 (1)	0.41 (1)	0.87 (1)	0.42 (1)
SVC	17 (90)	80 (339)	26 (81)	8 (10)	0.68 (0.9)	0.75 (0.81)	0.4 (0.53)	0.91 (0.97)	0.5 (0.66)
Gradient Boosting Classifier	7.26 (97.23)	99.72 (420)	6.28 (0)	17.74 (2.77)	0.29 (0.97)	0.94 (1)	0.54 (1)	0.85 (0.99)	0.38 0.99)

[a] The 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and evaluation are repeated 100 times to obtain mean scores. All scores pertain to mean values across 100 independent runs of each model. Test scores are shown in nonbrackets, while training scores are shown with brackets. Abbreviations: KNN, K Nearest Neighbors; SVC Support Vector Classifier, GBDT, Gradient Boosting Decision Tree; TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

Supplementary Table 4. CAN-BIND-1 external validation results (k = 100) with additional metrics[a].

	Mean Scores for Test and Train								
model	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	F1 score
Dummy_Classification	7.5 (23.87)	108.77 (424.29)	26.23 (101.71)	31.5 (101.13)	0.19 (0.19)	0.81 (0.81)	0.22 (0.19)	0.78 (0.81)	0.21 (0.19)
Random_Forest	22.11 (97.97)	109.97 (422.48)	25.03 (103.52)	16.89 (27.03)	0.57 (0.78)	0.81 (0.80)	0.47 (0.49)	0.87 (0.94)	0.51 (0.60)
Logistic_Regression	20 (106)	95 (399)	40 (127)	19 (19)	0.51 (0.85)	0.7 (0.76)	0.33 (0.45)	0.83 (0.95)	0.4 (0.59)
Elastic Net	19.97 (105.4)	94.89 (397.95)	40.11 (128.05)	19.03 (19.6)	0.51 (0.84)	0.7 (0.76)	0.33 (0.45)	0.83 (0.95)	0.4 (0.59)
KNearest_Neighbors	7 (125)	111(526)	24 (0)	32 (0)	0.18 (1)	0.82 (1)	0.23 (1)	0.78 (1)	0.2 (1)
Support_Vector_Machine	22 (110)	91 (420)	44 (106)	17 (15)	0.56 (0.88)	0.67 (0.80)	0.33 (0.51)	0.84 (0.97)	0.42 (0.65)
Gradient Boosting Classifier	13.05 (114.9)	123.83 (526)	11.17 (0)	25.95 (10.1)	0.33 (0.92)	0.92 (1)	0.54 (1)	0.83 (0.98)	0.41 (0.96)

[a] The 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and evaluation are repeated 100 times to obtain mean scores. All scores pertain to mean values across 100 independent runs of each model. Test scores are shown in nonbrackets, while training scores are shown with brackets. Abbreviations: KNN, K Nearest Neighbors; SVM Support Vector Machine, GBDT, Gradient Boosting Decision Tree; TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

Supplementary Table 5. STAR*D internal validation (k = 100) comparison of Balanced Accuracy P-values.[a]

	Dummy Classifier	Logistic Regression	Random Forest	Elastic Net	KNN	SVM	GBDT
Dummy Classification	1	4.74E-138*	7.23E-111*	8.98E-138*	2.20E-84*	9.28E-117*	5.31E-61*
Logistic Regression	4.74E-138*	1	2.95E-105*	3.14E-09*	0*	0*	1.20E-155*
Random Forest	7.23E-111*	2.95E-105*	1	2.35E-99*	1.08E-109*	0.070874258	1.84E-96*
Elastic Net	8.98E-138*	3.14E-09*	2.35E-99*	1	6.32E-210*	3.66E-149*	1.16E-150*
KNN	2.20E-84*	0*	1.08E-109*	6.32E-210*	1	0*	5.84E-33*
SVM	9.28E-117*	0*	0.070874258	3.66E-149*	0*	1	7.78E-116*
GBDT	5.31E-61*	1.20E-155*	1.84E-96*	1.16E-150*	5.84E-33*	7.78E-116*	1

*P < 0.05, with Bonferroni Correction (n= 49).

[a] The 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and internal evaluation are repeated 100 times to obtain mean balanced accuracy, and two-tailed t-tests were performed comparing model performance. Abbreviations: KNN, K Nearest Neighbors; SVM Support Vector Machine, GBDT, Gradient Boosting Decision Tree.

Supplementary Table 6. CAN-BIND-1 external validation (k = 100) comparison of Balanced Accuracy P-values.[a]

	Dummy Classifier	Logistic Regression	Random Forest	Elastic Net	KNN	SVM	GBDT

Dummy Classification	1	2.54E-72*	2.81E-112*	1.27E-71*	0.628499721	5.47E-79*	1.76E-78*
Logistic Regression	2.54E-72*	1	1.17E-159*	0.006315537	0*	0*	2.17E-27*
Random Forest	2.81E-112*	1.17E-159*	1	8.49E-157*	1.85E-230*	6.47E-148*	8.30E-94*
Elastic Net	1.27E-71*	0.006315537	8.49E-157*	1	1.55E-283*	9.68E-98*	2.69E-28*
KNN	0.628499721	0*	1.85E-230*	1.55E-283*	1	0*	2.15E-162*
SVM	5.47E-79*	0*	6.47E-148*	9.68E-98*	0*	1	1.84E-06*
GBDT	1.76E-78*	2.17E-27*	8.30E-94*	2.69E-28*	2.15E-162*	1.84E-06*	1

*P < 0.05, with Bonferroni Correction (n= 49).
[a] The 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and external evaluation are repeated 100 times to obtain mean balanced accuracy, and two-tailed t-tests compared model performance. Abbreviations: KNN, K Nearest Neighbors; SVM Support Vector Machine, GBDT, Gradient Boosting Decision Tree.

Supplementary Table 7. P-values for mean balanced accuracy, comparing STAR*D internal validation (k = 100) vs CAN-BIND-1 external (k = 100) validation performance.[a]

	Dummy Classification	Logistic Regression	Random Forest	Elastic Net	KNN	SVC	GBDT
P-value	0.291463749	0*	4.73E-30*	1.28E-225*	0*	0*	3.20E-05*

*P < 0.05, with Bonferroni Correction (n= 7).
[a] The 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and evaluation are repeated 100 times to obtain mean balanced accuracy. Two tailed t-tests were performed, comparing internal vs external model performance. Abbreviations: KNN, K Nearest Neighbors; SVM Support Vector Machine, GBDT, Gradient Boosting Decision Tree.

Supplementary Table 8: Balanced accuracy and AUC score of various machine learning models on the STAR*D holdout set including when using Elastic Net feature selection.[a]

Models	Mean Balanced Accuracy	Mean AUC Score
Logistic Regression (k = 480)	72%	0.80
Logistic Regression (ENet k = 61)	66%	0.76
Elastic Net Regression (k = 480)	68%	0.75
Random Forest (k = 480)	72%	0.78
Random Forest (ENet k = 61)	70%	0.79
SVC (k= 480)	71%	0.79
SVC (ENet k = 61)	68%	0.75
GBDT (k = 480)	62%	0.77
GBDT (ENet k = 61)	65%	0.76
KNN (k = 480)	54%	0.76
KNN (ENet k = 61)	51%	0.62

[a] Mean performance from 100 runs of each model is assessed using all 480 features and a reduced feature set of 61 features, via cross-validated elastic net (ENet). N represents the number of features that a model was trained and evaluated on. Abbreviations: SVC, Support Vector Classifier; GBDT, Gradient Boosting Decision Tree; KNN, K Nearest Neighbors.

Supplementary Table 9. Balanced accuracy and AUC scores (brackets) across different machine learning models after Elastic Net feature selection, assessed on holdout sets.[a]

	Mean balanced accuracies (mean AUC scores)	
Model	STAR*D Holdout Set (ENet k = 30)	CAN-BIND-1 External Validation (ENet k = 30)
Logistic Regression	72% (0.82)	60% (0.69)
Random Forest	71% (0.81)	65% (0.74)
SVC	72% (0.78)	59% (0.66)
GBDT	60% (0.73)	59% (0.67)
KNN	59% (0.71)	62% (0.68)

[a] The 30 features were derived via Elastic Net Selection from the 100 features (k) found in both CAN-BIND and STAR*D are used for these evaluations. Training and evaluation are repeated 100 times to obtain mean scores. Abbreviations: AUC, area under the curve; SVC, Support Vector Classifier; GBDT, Gradient Boosting Decision Tree; KNN, K Nearest Neighbors.

Supplementary Table 10. P-values for Random Forest mean balanced accuracy on STAR*D internal validation, across several feature set variations.[a]

	Full (k=480)	No QIDS-SR or Q-LES-Q-SF Full STAR*D (k=391)	Overlapping (k = 100)	QIDS-SR only (k=47)	Q-LES-Q-SF only (k=16)	QIDS-SR + Q-LES-Q-SF (k=63)	No QIDS-SR or Q-LES-Q-SF Over-lapping (k=37)
Full (k =480)[b]	1	9.77E-97	0.008102091	2.74E-39	1.50E-82	0.011822109	2.24E-76
No QIDS-SR or Q-LES-Q-SF Full STAR*D (k =391)[c]	9.77E-97	1	8.71E-82	5.85E-47	7.13E-37	6.44E-92	0.000167881
Overlapping (k = 100)[d]	0.008102091	8.71E-82	1	2.13E-26	1.19E-59	0.567348003	4.80E-64
QIDS-SR only (k = 47)[e]	2.74E-39	5.85E-47	2.13E-26	1	1.44E-12	8.10E-33	5.75E-30
Q-LES-Q-SF only (k =16)[f]	1.50E-82	7.13E-37	1.19E-59	1.44E-12	1	7.44E-75	4.65E-17
QIDS-SR + Q-LES-Q-SF (k = 63)[g]	0.011822109	6.44E-92	0.567348003	8.10E-33	7.44E-75	1	1.25E-71
No QIDS-SR or Q-LES-Q-SF Overlapping (k = 37)[h]	2.24E-76	0.000167881	4.80E-64	5.75E-30	4.65E-17	1.25E-71	1

[a] Training and evaluation are repeated 100 times to obtain mean scores. [b] Full 481 features. [c] Full feature set, excluding QIDS-SR or Q-LES-Q related features. [d] Overlapping 100 features. [e] Only features related to QIDS. [f] Only features related to QLESQ-SF. [g] Both QIDS-related and QLESQ-SF-related features. [h] All overlapping features, excluding those related to QIDS-SR or QLESQ-SF. Abbreviations: Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.

Supplementary Table 11. A comparison of Random Forest balanced accuracies and AUC scores for top performing models on STAR*D holdout set, using several feature variations[a].

	Mean balanced accuracies (mean AUC scores)			
	Elastic Net	Logistic Regression	Random Forest	Support Vector Classifier
Full (k= 480)[b]	0.68 (0.75)	0.72 (0.80)	0.72 (0.78)	0.71 (0.79)
QIDS-SR (k = 72)[c]	0.69 (0.77)	0.63 (0.76)	0.69 (0.76)	0.73 (0.77)
QLESQ-SF (k = 17)[d]	0.73 (0.77)	0.68 (0.75)	0.66 (0.78)	0.71 (0.77)
QIDS-SR + QLESQ-SF (k = 89)[e]	0.73 (0.82)	0.73 (0.82)	0.69 (0.78)	0.73 (0.81)
Exclude QIDS-SR + QLESQ-SF (k = 391)[f]	0.66 (0.68)	0.57 (0.70)	0.64 (0.72)	0.66 (0.74)

[a] Training and evaluation are repeated 100 times to obtain mean scores. [b] Full 481 features. [c] Only features related to QIDS. [d] Only features related to QLESQ-SF. [e] Both QIDS and QLESQ-SF features. [f] All features, excluding those related to QIDS-SR or QLESQ-SF. Abbreviations: AUC, area under the curve; k, number of features, Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.

Supplementary Table 12. P-values for Random Forest mean balanced accuracy on CAN-BIND-1 (k=100) external validation, across several feature set variations.[a]

	Overlapping (k=100)	QIDS-SR only (k=47)	Q-LES-Q-SF only (k=16)	QIDS-SR + Q-LES-Q-SF (k=63)	No QIDS-SR or Q-LES-Q-SF Overlapping (k=37)
Overlapping (k = 100)[b]	1	1.99E-51	3.24E-171	3.82E-11	2.09E-142
QIDS-SR only (k = 47)[c]	1.99E-51	1	7.85E-150	3.40E-31	1.81E-124

Q-LES-Q-SF only (k =16)[d]	3.24E-171	7.85E-150	1	1.27E-165	2.29E-13
QIDS-SR + Q-LES-Q-SF (k = 63)e	3.82E-11	3.40E-31	1.27E-165	1	3.32E-137
No QIDS-SR or Q-LES-Q-SF Overlapping (k = 37)[f]	2.09E-142	1.81E-124	2.29E-13	3.32E-137	1

[a] Training and evaluation are repeated 100 times to obtain mean scores. [b] Overlapping features compatible between STAR*D and CAN-BIND-1. [c] Only features related to QIDS. [d] Only features related to QLESQ-SF. [e] Both QIDS-related and QLESQ-SF-related features. [f] All features excluding those related to QIDS-SR or QLESQ-SF. Abbreviations: Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.

Supplementary Table 13. P-values for Random Forest balanced accuracy, comparing STAR*D internal validation vs CAN-BIND-1 external validation performance, across feature sets.[a]

Feature Set:	Overlapping (k = 100)	QIDS-SR (k = 47)	Q-LES-Q-SF (k=16)	QIDS-SR + Q-LES-Q-SF (k=63)	No QIDS-SR or Q-LES-Q-SF (k=37)
P-Value:	4.73E-30*	1.69E-27*	3.40E-157*	1.11E-60*	9.45E-97*

*P < 0.05, with Bonferroni Correction (n= 5).

[a] Training and evaluation are repeated 100 times to obtain mean scores. [b] Overlapping features. [c] Only features related to QIDS. [d] Only features related to QLESQ-SF. [e] Both QIDS-related and QLESQ-SF-related features. [f] All features excluding those related to QIDS-SR or QLESQ-SF. Abbreviations: Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.

Appendix 1: Description of the overlapping features used in this work.

One hundred features overlapped appropriately, between the STAR*D and CAN-BIND-1 datasets. Each STAR*D variable included in the overlapped dataset that was used for both internal and external validation, is described below along with its CAN-BIND-1 equivalent scale. Overlapping Features (k = 100):

Demographics:	Educat - # of years in formal education; Empl – Current employment status; Totincom – Monthly household income; Marital – current marital status; Interview age - Age in months at the time of the interview/test/sampling/imaging; Gender - Female, Male.
Work Productivity and Impairment (WPAI) (CAN-BIND-1 equivalent –Lam Employment Absence and Productivity Scale (LEAPS))	Totalhrs – hours missed + hours worked; Wpai02 – work hours missed due to health problems.
Psychiatric Hx (CAN-BIND-1 equivalent – Mini International Neuropsychiatric Interview (MINI))	Dep – family history of depression; Dage – age of onset of first major depressive episode (MDE); Episode_date – onset of current MDE; Epino – number of MDEs; Pd_ag – Panic w/ agoraphobia; Pd_noag – panic w/ agoraphobia; Alcoh – alcohol abuse; Anorexia – anorexia present; Antis – antisocial personality; Bulimia – current diagnosis of bulimia; Gad_phx – generalized anxiety; Ocd_phx - obsessive-compulsive; Psd – post traumatic stress; Amphet – amphetamine abuse; Soc_phob – social phobia; Imput_Any anxiety – phx01_psd + pd_noag + pd_ag+ soc_phob + gad_phx + specphob; PHYHIS_MDD_PREV:: - number of MDE's >= 2 (similar to epino).
QIDS-SR (W0sr = Week 0 Self-rated, W2sr = Week2 self-rated)	W0sr_qstot – QIDS Total Score; W0sr_vcntr – Concentration/Decision Making; W0sr_vvwsf – self outlook; W0sr_vsuic – suicidal ideation; W0sr_vintr – Involvement in interests/activities; W0sr_vengy – Energy/Fatigability; W0sr_vslow – psychomotor slowing; W0sr_vagit – psychomotor agitation; W0sr_vsoin – sleep onset insomnia; W0sr_vmnin -mid-nocturnal insomnia; W0sr_vemin – early morning insomnia; W0sr_vhysm - hypersomnia; W0sr_vmdsd – Sad Mood; W0sr_vapdc – decreased appetite; W0sr_vapin – increased appetite; W0sr_vwtdc – decreased weight in past 2 weeks; W0sr_vwtin – increased weight in past 2 weeks.
QIDS Atypical Baseline QIDS Atypical Week 2	Imput_QIDS_SR_appetite_domain_week0 – including both increased and decreased appetite; Imput_QIDS_SR_appetite_domain_week2 – including both increased and decreased appetite; Imput_QIDS_SR_ insomnia_week0 – including sleep-onset, mid-nocturnal, and early-morning subtypes; Imput_QIDS_SR_ insomnia_week2 – including sleep-onset, mid-nocturnal, and early-morning subtypes; Imput_QIDS_SR_ overeating_week0 - including increased appetite and increased weight; Imput_QIDS_SR_ overeating_week2- including increased appetite and increased weight; Imput_QIDS_SR_ psychomotor_week0 - including psychomotor slowing and agitation; Imput_QIDS_SR_ psychomotor_week2 - including psychomotor slowing and agitation; Imput_QIDS_SR_ sleep_week0 - including all 3 subtypes of insomnia + hypersomnia; Imput_QIDS_SR_ sleep_week2 - including all 3 subtypes of insomnia + hypersomnia; Imput_QIDS_perc_change – percent change in total score from week 0 to week 2.
QLESQ Total	Total QLESQ baseline score; Qlesq01 – Overall physical Health; Qlesq02 - Mood; Qlesq03 – Work Performance; Qlesq04 – Household Activities; Qlesq05 – Social Relationships; Qlesq06 – Familial relationships; Qlesq07 – Leisure Activities; Qlesq08 – Ability to function; Qlesq09 – Sexual drive; Qlesq10 – economic status; Qlesq11 – Housing situation; Qlesq12 – Physical Mobility; Qlesq13 - Vision; Qlesq14 -Overall well-being; Qlesq15 - Treatment; Qlesq16 – overall contentment.
Work and Social Adjustment Scale (CAN-BIND-1 Equivalent – Sheehan Disability Scale, SDS))	Wsas01 – Work is impaired; Wsas02 – Home management impairment; Wsas03 – Social activities impairment;

Some models are trained with all 480 available STAR*D features which are well defined elsewhere, and are accessible on the NDA data archive.

Appendix 2: Participant Selection

Aside from clinical inclusion/exclusion criteria included in STAR*D and CAN-BIND-1 studies, several data processing steps were undertaken that affected subject selection, resulting in a total of 651 examples for STAR*D and 178 for CAN-BIND-1. Due to difference in the formatting of the STAR*D and CAN-BIND-1 datasets, different processing steps were applied and are summarized below, along with rationale as necessary.The STAR*D selection steps and rationale are described below and can found in classes.py and stard_preprocessing_manager.py. They are listed in the order they were performed in the script, and the number of participants excluded via each step is shown.

Starting number of unique participants, after initial dataset cleaning: 3818

1. Remove participants who went into early follow-up or Level 2, likely due to adverse side effects of the medication (3739) (-79)
2. Only include participants who have at least one Q-LES-Q-SF value between 4-9 weeks of the study, using last-outcome-carried-forward (LOCF) (3657) (-82)
3. Drop rows with missing Q-LES-Q-SF values. (3592) (-65)
4. Remove duplicate rows. (3592) (-0)
5. Remove rows involving Level 3 or Level 4, as the scope is limited to the end point of Level 1 in the STAR*D trial. If a Q-LES-Q-SF value located between 4-9 weeks but was associated with a first occurrence of Level 2 for that patient, it was kept. The rationale is that the patient has just stopped their Level 1 drug. (3592) (-0)
6. Remove participants where there is one or less QoL data point. (-2879)
7. Remove participant if their baseline Q-LES-Q-SF value is beyond the 4th week of the study (-7)
8. Remove participant if their final Q-LES-Q-SF is before the 4th week of the study (- 38)
9. Remove patients who start with a Q-LES-Q-SF baseline score that is already above the established threshold for a nonimpaired QoL (≥67). (--17)

Final number of unique participants: 651

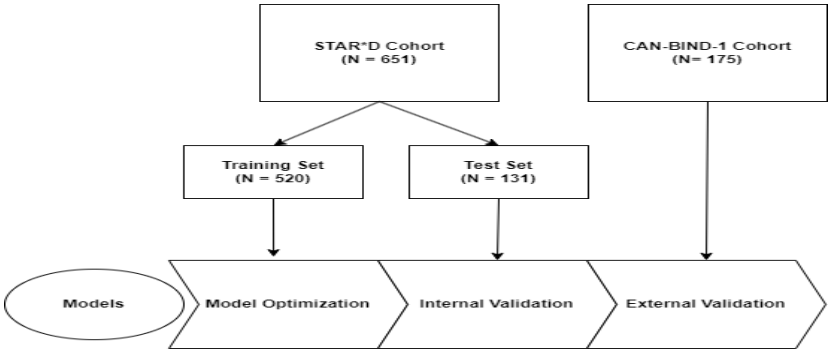
The CAN-BIND-1 selection steps and rationale are described below and can found in canbind_ygen.py:

Starting number of unique participants, after initial dataset cleaning: 323

1. Only keep data from the treatment group, excluding the control group (211) (-112)
2. Remove rows with missing values for Q-LES-Q-SF (-0)
3. Only retain participants that have both a baseline + Week 8 Q-LES-Q-SF score (176) (-35)
4. Exclude participants if missing ≥ 5 questions from the 14 items Q-LES-Q-SF scale (-0)
5. Remove patients who start with a Q-LES-Q-SF baseline score that is already above the established threshold for a nonimpaired QoL (67). (-1)

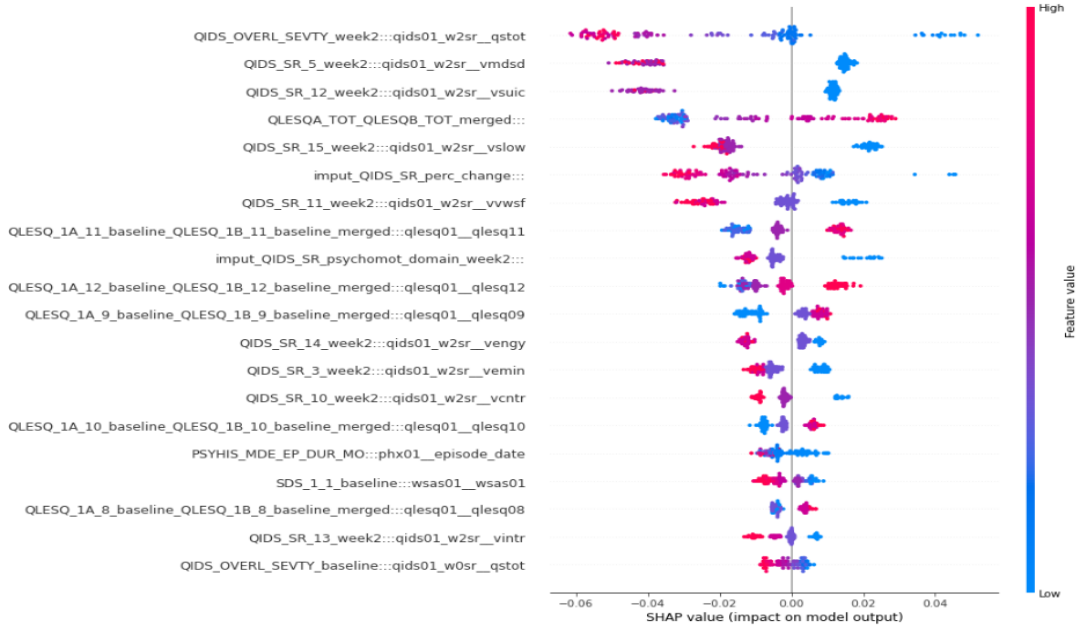
Final number of unique participants: 175

Supplementary Figure 1. A flowchart summary of model training and evaluation, where N represents sample size for each of the steps.



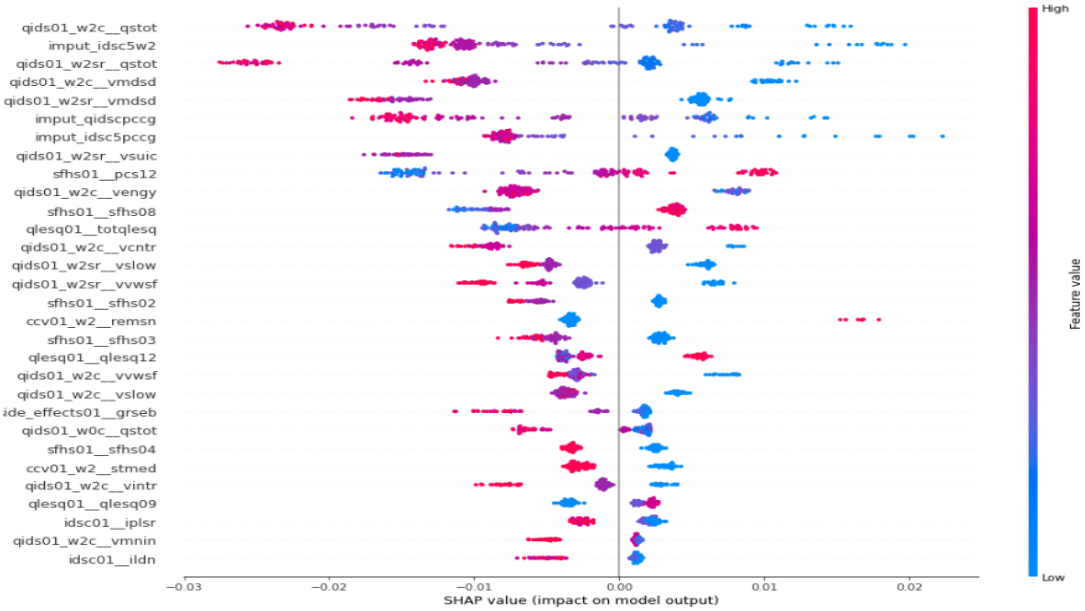
Abbreviations: CANBIND-1, Canadian Biomarker Integration Network in Depression; STAR*D, Sequenced Treatment Alternatives to Relieve Depression.

Supplementary Figure 2. SHAP values taken from a randomly sampled Random Forest model, trained on the overlapping feature set (k = 100).



Abbreviations: SHAP, SHapley Additive exPlanations; CANBIND-1, Canadian Biomarker Integration Network in Depression; STAR*D, Sequenced Treatment Alternatives to Relieve Depression; Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.

Supplementary Figure 3. SHAP values taken from a randomly sampled Random Forest model, trained on the full dataset (k= 480).



Abbreviations: SHAP SHapley Additive exPlanations; CANBIND-1, Canadian Biomarker Integration Network in Depression; STAR*D, Sequenced Treatment Alternatives to Relieve Depression; Q-LES-Q-SF, Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form; QIDS-SR Quick Inventory of Depressive Symptomatology - Self Report.