Supplementary Materials

Overview:

- Data access information
- Supplementary Methods
    - Notes on sample
    - CONSORT-style flow chart
    - Predictor variables
    - Machine learning methods
        - Hyperparameter selection
        - Performance metrics
        - Imbalanced class considerations
- Sensitivity Analyses using alternative exclusion criteria
    - SA1. Demographic only model
    - SA2. Exclusion of participants for missing all DSM data
    - SA3. Exclusion of participants for missing DSM data and responses that may be internally-inconsistent
- Detailed performance metrics
    - Permutation metrics for Table 2 in main manuscript
    - Group-level variable importance plots for models predicting reasons for non-initiation
- Breakdowns of outcomes by age, race, and gender
- R package citations
- Complete variable list
- References

Data access information.

Data were downloaded from the Inter-university Consortium for Political and Social Research (ICPSR), and are available upon request from https://www.icpsr.umich.edu/icpsrweb/index.jsp. Institutional review board approval and informed consent were not needed because this was a secondary analysis of data from a public use file.

Supplementary Methods

Sample notes

To minimize issues of response bias, the sample was restricted to individuals for whom a diagnosis was issued in the previous 12 months (i.e. to ensure that participants are recalling an event that was at least in the previous 12 months). Thus, it is possible that people who were depressed in that period but who received their diagnosis previously were not included (assuming that those people considered themselves as not having a diagnosis of depression in that period).

Conversely, although all members of the sample had received a doctor's diagnosis of depression in the past 12 months, only about half of the sample actually met criteria for current MDE (see main manuscript). Thus, it is possible that some individuals in the sample had subthreshold symptom levels, or perhaps a depressive episode in the last year that has since resolved. It is not clear how these individuals might have affected the predictability of treatment initiation. In addition, it is known that at least some patients with major depression experience sudden therapeutic gains and may have recovered without treatment[1,2], although long-term outcomes are generally not favorable for untreated patients[3].

Predictor variables.

We included the following socio-demographic information: age (five categories: 18-25, 26-34, 35-49, 50-64, 65+), sex, race (White, Black/African American, Native American/Alaskan, Native Hawaiian/Pacific Islander, Asian, Multi-racial, Hispanic), marital status (married, widowed, divorced/separated, never married), the number of children <18 in the household, level of education, county type (large metro, small metro, non-metro), family household income (seven categories in thousands of dollars: less than 10, 10-19.99, 20-29.99, 30-39.99, 40-49.99, 50-74.99, and 75 or more), and separate binary indicators of whether the individual was covered by Medicare, Medicaid/CHIP, or by private health insurance. Relating to behavioral health, we included the nine major depression items from the DSM-IV(29) (in other words, a 0-1 coded version of the PHQ-9), six items from the Kessler-6 (30) (a psychological distress scale), and self-reported endorsement of having thought about, planned, or attempted suicide before. Finally, we included a brief medical history (selected by convenience): how many of the last 30 days they smoked cigarettes, the number of times they were in the emergency room in the last year, their self-reported overall health status (1 = "Excellent" to 5 = "Poor"), and whether they have ever been diagnosed with the following health conditions in their life: anxiety disorder, asthma, bronchitis, liver cirrhosis, diabetes, heart disease, hepatitis, high blood pressure, pneumonia, STD, sinusitis, sleep apnea, stroke, or ulcers.

Machine learning methods

Algorithm selection. We used a tree-based machine learning algorithm called XGBoost (Extreme Gradient Boosting, http://xgboost.readthedocs.io/en/latest/model.html). The XGBoost algorithm is based on the original gradient boosting machine algorithm[4,5] but includes a number of optimizations for speed and performance based on open-source development in recent years. Rather than fitting one strong model (i.e. a deep or complex tree) to a dataset, a gradient boosting machine is built by combining several weakly predictive models to relate the predictors and outcome[4]. Crucially, when each successive tree is fit, the model focuses on the data that previous models failed to predict[6]. The xgboost algorithm includes several features to help minimize overfitting, including subsampling of features (i.e. each tree is built using just a subset of all predictor variables) and subsampling of observations (i.e. only a subset of patients are used to build each tree), and applying regularization over the ensemble of trees.

Supplementary CONSORT-style flow diagram illustrating samples used for each analysis

1) Participants from annual national survey on drug use and mental health with self-reported diagnosis of depression

2008 N = 55,110
2009 N = 55,234
2010 N = 57,313
2011 N = 58,397
2012 N = 55,268
2013 N = 55,160
2014 N = 55,271

N = 391,753

Excluded:
- Children <18 years old (n = 121,526)
- No diagnosis of MDD in last 12 months (n = 249,398)

n = 20,829

2) People who say they needed treatment

Excluded:
- No outcome response (n = 44)

n = 20,785

(30.2% endorse needing but not receiving)

3) If they did not get treatment, why not?

Excluded:
- Did not indicate unmet neet (n = 14,514)
- No reasons given for not getting treatment (n = 61)

n = 6,210

Self-reported reasons for not getting MH treatment:

47.7% Couldn't afford cost
22.2% Thought they could manage without treatment
16.7% Didn't know where to go
15.3% Some other reason
15.2% Thought they might be committed or forced to take meds
14.2% Didn't have time
11.7% Not enough health insurance coverage
11.0% Concerned about neighbors' opinion
11.0% Didn't think treatment would help
9.7% Concerned about confidentiality
8.6% Didn't think they needed it
8.1% Concerned about effect on job
6.5% Health insurance didn't cover it
6.5% Didn't want others to find out
5.8% Had no transportation or treatment too far

Individual-subject variable importance. Although machine learning approaches are often more accurate than traditional statistical approaches, their increased complexity can result in lower interpretability. This has been characterized as the "black box" nature of machine learning models. Furthermore, although predictions are made on an individual subject level, researchers typically interpret models using group-level variable importance measures (e.g. how much reduction in error is seen when a variable is included, or the gain contribution of each feature to the model, or for parametric models, the coefficient assigned to a given variable). Although these approaches do give some insight into which variable is the most important across all predictions, there is no guarantee that it is the most important variable for a particular individual.

With this in mind, we have developed and introduced an open-source software library for deriving individual-subject level variable importance measures from an xgboost ensemble (https://github.com/AppliedDataSciencePartners/xgboostExplainer). The input for the library is a trained xgboost ensemble, and an individual subject for whom we need to interpret the prediction of the ensemble. The output of the library is a figure that illustrates the impact of each feature for the particular individual, alongside the overall prediction of the xgboost algorithm. Specifically, the library calculates the change in the log-odds prediction at each branch point in the route taken by the individual subject through the ensemble of trees and attributes the change in prediction to the feature at the branch. Summing the individual contributions for each feature produces the overall feature impact, which is unique to the observation. The intercept term is simply the log-odds prediction at the root of the first tree (i.e. before any branching has taken place)

This methodology has the crucial property that the sum of the intercept and feature impacts exactly equals the overall log-odds prediction. It is critical to note here that these "impacts" are not static coefficients as in a logistic regression— the impact of a feature is dependent on the specific path that the observation took through the ensemble of trees.

The explainer figure is laid out as follows. The x-axis represents the probability of the response variable, which in this case refers to the probability that a patient will fail to receive needed treatment for mental health. Values to the far right i.e. high probabilities thus indicate a high probability that a patient will fail to initiate treatment. At the top of the figure is the overall predicted probability of the model for that individual. The rest of the bars from the bottom to the top illustrate a waterfall decomposition of each feature's contribution to the overall predicted probability. The bars are in rank order of the size of their contribution and, for convenience, effects with log-odds contributions that are smaller than a certain threshold (e.g. 0.01) are aggregated into a single bar for "others". The value inside the bar is the change in log-odds attributable to that individual feature at that specific level (e.g. gender equal to female, age equal to 25 years, etc).

Hyperparameter selection. We selected appropriate hyperparameters using a grid search over a pre-determined range of reasonable parameters using a AUC-optimisation process. The number of rounds of boosting was chosen from a small range ($nrounds \in \{100, 250, 500\}$). The maximum depth of each tree was small ($max\_depth \in \{1, 2\}$). We included a step size shrinkage to prevent overfitting, ($eta \in \{0.05, 0.1\}$) and varied the minimum loss reduction required to make a further partition on a leaf node of the tree ($gamma \in \{0, 0.1, 0.2\}$). The minimum child weight was fixed at 1, the subsample rate was fixed at 50% (i.e. half of the training data within each fold was available for each tree), the column sample by tree was fixed at 50% (i.e. 50% of predictors were available for building each tree). Further information about these parameters is available in the *xgboost* documentation, http://xgboost.readthedocs.io/en/latest/parameter.html.

Performance metrics. Some performance measures (e.g. accuracy) are inappropriate when the outcomes are not balanced. For instance, if you are predicting an outcome like credit card fraud (which only occurs in 5% of individuals), then a trivial algorithm that predicts "no fraud" all the time will have an accuracy of 95% (it will be correct all the time there is no fraud, and incorrect whenever there is fraud). Therefore, we focused on balanced accuracy, calculated as the arithmetic mean of the model's sensitivity and specificity. This metric is centered on 50%, as evidenced in our permutation tests of these data, and thus is a more informative metric than simple accuracy in this context. To understand how much more likely a positive test result is in a participant with the non-desired outcome (e.g. not coming back for treatment, or endorsing a specific reason for not getting treatment), we also calculated Positive Likelihood Ratios [calculated as sensitivity/(1-specificity)], as in[7].

Permutation testing. The statistical significance of each model was assessed using a permutation test: outcomes were shuffled before the modeling pipeline was applied (200 repeats), and the unshuffled performance metric was compared to the distribution of shuffled-metrics ($\alpha = 0.05$, one-sided test).

Imbalanced class proportions. Machine-learning approaches applied to data with severe class imbalances often produce algorithms that do not accurately predict the minority class, and predictions are biased towards the majority class. This problem was salient in evaluating the individual reasons why a participant did not get treatment: in the training data, for each possible reason, as few as 6.1% of participants may have endorsed a specific reason (e.g. "had no transport or treatment too far"). We took two steps to counteract class imbalances when predicting outcomes in this study. First, in the training data, we used bootstrapped up-sampling, i.e. we randomly sampled (with replacement) the minority class to be the same size as the majority class. **Test-fold and validation data were not upsampled, and thus remained representative of the true imbalanced classes**. Second, we used an adjusted probability threshold before applying models to the validation data: for each reason, we calculated the rate of endorsement in the training data (R), and then only gave positive predictions to individuals for whom the probabilistic output of the classifier was in the top R fraction of predicted probabilities. For example, lack of transport (5.6% training endorsement): a positive prediction was only given to participants who had output probabilities in the top 5.6% of subjects. Note that this is likely to give suboptimal performance relative to typical calibration procedures (these would usually consider the test prevalence, which is 7.2% in the case of lacking transport), but has better external validity since it does not involve any insight drawn from the validation sample.

Alternative algorithms. Machine learning algorithms are particularly useful in cases where there are a large number of predictor variables. This was not the case here, and indeed performance for our main case finding model was similar when we compared it to a generalized linear model or an elastic net regression (i.e. penalized logistic regression). In this study, both the elastic net regression and the xgboost model used fewer variables than traditional regression methods in attaining this performance, and using fewer variables would in turn reduce the time taken to collect the data required from a patient to use the model.

Sensitivity Analyses

We conducted a number of sensitivity analyses to better understand the impact of experimenter degrees of freedom in these analyses.

SA1. Demographic-only model

We examined whether it was really necessary to use all these variables to predict non-engagement, or whether it could have been done with sociodemographic variables alone. For this analysis, we used the same sample selection criteria as the main analysis (Figure 1, part 2), i.e. we trained on 17,325 individuals from the 2008-2013 cohort, and tested on 3,460 individuals from the 2014 cohort. However, in this analysis, we only included the following predictor variables: age (five categories: 18-25, 26-34, 35-49, 50-64, 65+), sex, race (White, Black/African American, Native American/Alaskan, Native Hawaiian/Pacific Islander, Asian, Multi-racial, Hispanic), marital status (married, widowed, divorced/separated, never married), the number of children <18 in the household, level of education, county type (large metro, small metro, non-metro), family household income (seven categories in thousands of dollars: less than 10, 10-19.99, 20-29.99, 30-39.99, 40-49.99, 50-74.99, and 75 or more), and separate binary indicators of whether the individual was covered by Medicare, Medicaid/CHIP, or by private health insurance.

Predictive performance was greatly reduced both during cross-validation and in the external validation sample. In the external validation set, the model BAC was 61.7% which compares poorly with the performance of the model that also included physical and behavioral health variables. This suggests that meaningful information was extracted beyond simple sociodemographic associations.

Supplementary Table 1: Cross-validated performance metrics for predicting engagement using a sociodemographic-only model

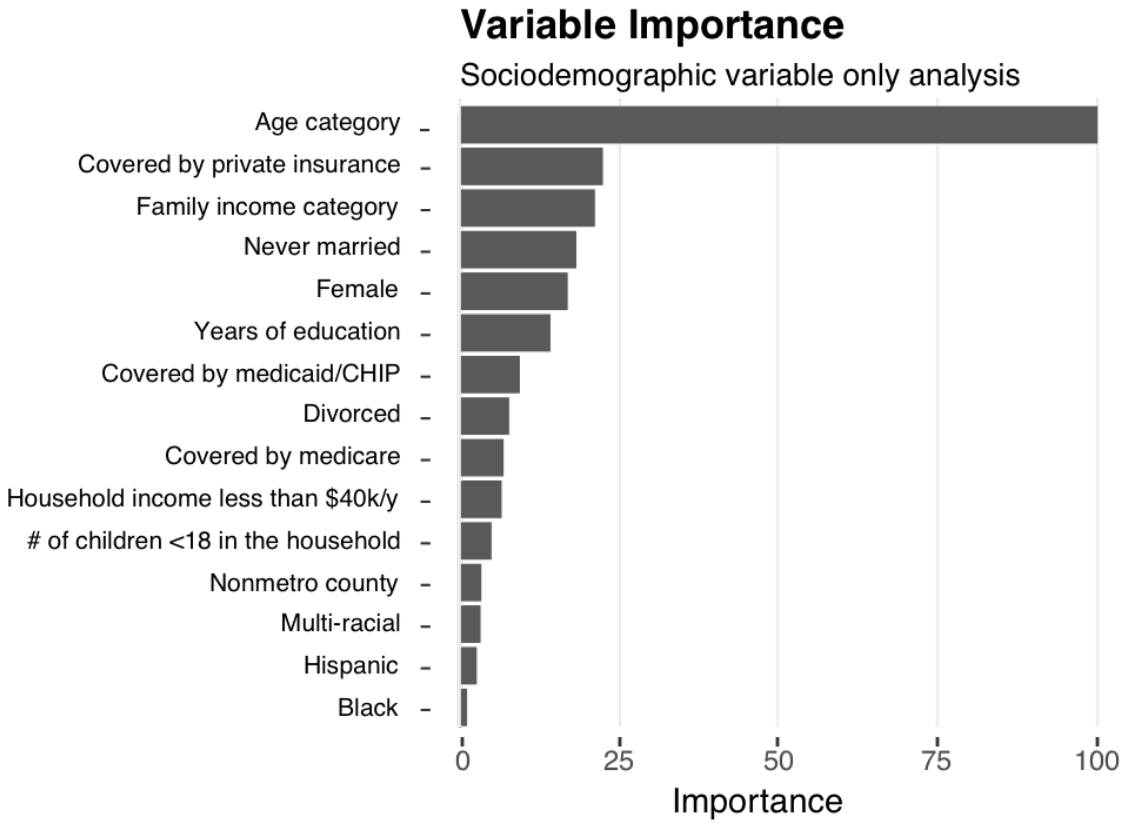
ST1	Mean	SD
AUC	0.630	0.013
Accuracy	0.557	0.010
Kappa	0.152	0.019
Sensitivity	0.686	0.021
Specificity	0.501	0.013
PPV	0.377	0.009
NPV	0.784	0.012
Balanced Accuracy	0.593	0.011

Supplementary Table 2: Performance metrics based on external validation of sociodemographic-only model

ST2	
Accuracy	0.6026 95% CI [0.5861, 0.619]
Kappa	0.192
Sensitivity	0.650
Specificity	0.584
PPV	0.380
NPV	0.809
Balanced Accuracy	0.617

Supplementary Figure 1: Variable importance for sociodemographic only model.

We measured variable importance as the average improvement in accuracy (i.e. Gain) brought by a particular variable when it is used. For plotting purposes, all variable importances were scaled as a percentage of the largest variable importance (i.e. the most influential variable was set to 100% and all others are relative to this).



SA2. Exclude subjects who were missing all 9 DSM items

For most predictor variables, there was very little missing data (less than 1%). However, some participants (8,494, or 41% of the 20,785) were missing all 9 of the DSM items. In the main analysis, we retained these subjects but used categorical imputation for missing data (imputing "-1" whenever an item was missing). We did not want to exclude participants who were missing DSM items because these participants were statistically significantly different at the group level from participants who were not missing these items according to a number of variables. Overall these patients who did not have any DSM items appeared to have fewer psychiatric concerns, although in a large sample size a statistically significant different may be small in practice. For example, participants who had no DSM items available were: more likely to engage in treatment; less likely to endorse suicidal ideation, suicide plans, or suicide attempts; older; less likely to have a history of anxiety; and endorsed less severe responses on the Kessler-6 scale. Therefore, we conducted a sensitivity analysis that excluded participants who were missing all 9 DSM items to ensure that the categorical imputation and the inclusion of these participants did not drive the predictability of the dependent variables.

In this analysis, we included 12,291 participants (i.e. 20785 (main analysis) - 8494 (missing DSM items). This included 10,195 participants in the 2008-2013 training set, and 2,096 participants in the 2014 external validation set. Amongst the training set, 39.2% of participants endorsed needing treatment but not getting it. In the validation set, 35.9% of participants endorsed needing treatment but not getting it. All statistical analyses were conducted as described in the main manuscript.

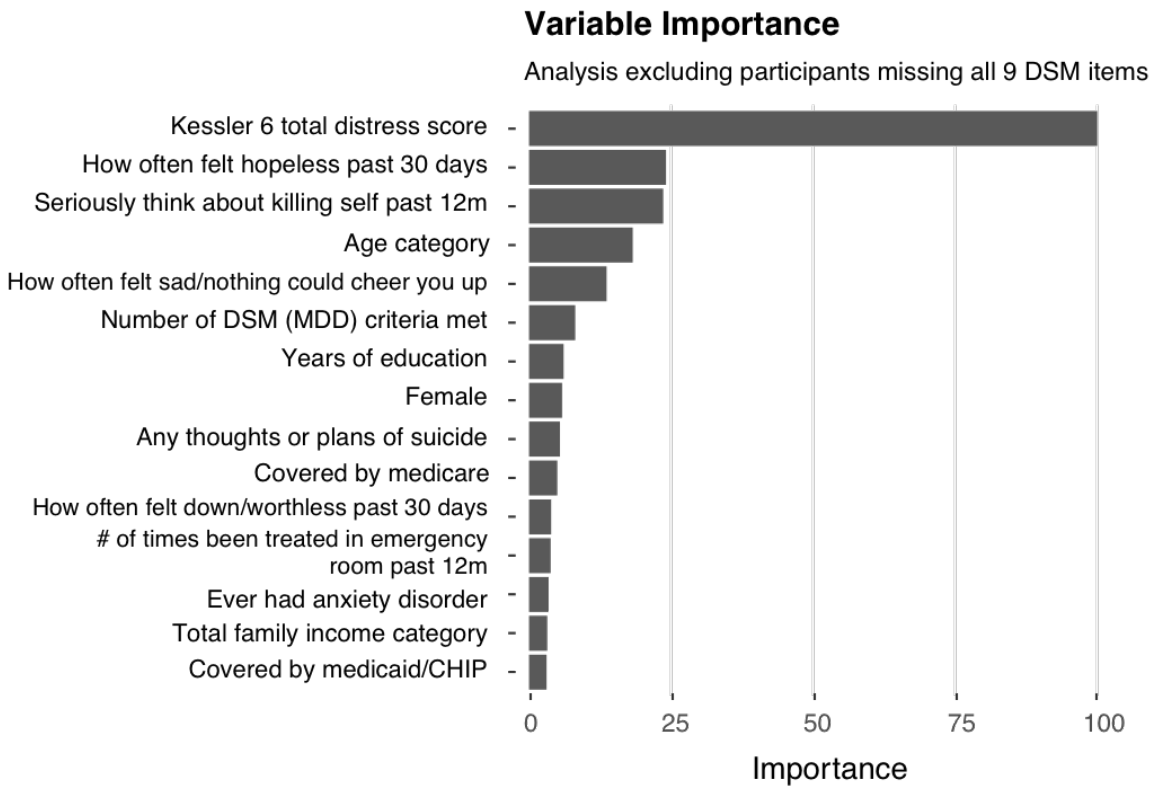
Once again, during cross-validation within the 2008-2013 data, model performance was above chance and broadly comparable to the performance metrics obtained in the main analysis. For example, balanced accuracy in this sample was 67.6% on average across folds and repeats (SD = 1.6%), compared to 70.5% obtained in the larger sample that included the subjects with missing data.

Supplementary Table 3: Cross-validated performance metrics for predicting engagement when participants were excluded if they were missing all DSM items.

ST3	Mean	SD
AUC	0.742	0.016
Accuracy	0.668	0.017
Kappa	0.336	0.031
Sensitivity	0.714	0.021
Specificity	0.638	0.024
PPV	0.560	0.018
NPV	0.776	0.014
Balanced Accuracy	0.676	0.016

We inspected the average improvement in accuracy (i.e. Gain) brought by a particular variable when it is used (variable importance) to see whether the predictors identified in this sample differed to those observed in the primary analysis. Nine of the top 10 coefficients in this analysis were also in the top 10 variables for the primary analysis.

Supplementary Figure 2: Variable importance in predicting engagement amongst participants who were not missing all 9 DSM items.



This model successfully generalized to predict engagement in the external validation set, with performance comparable (but numerically lower) than the performance observed in the primary analysis. For instance, the balanced accuracy in the external validation sample here (67.7%) was marginally lower than for the main analysis (70.5%). Nonetheless, balanced accuracy was significantly better than chance, mean = 0.677, (exact binomial) 95% confidence interval [0.656, 0.697], $p<0.0001$. This suggests that the imputation of large numbers of DSM items for some participants in the main analysis did not drive the ability of the model to predict our outcome of interest.

Supplementary Table 4: Performance metrics based on external validation of analysis excluding participants missing all 9 DSM items.

ST4	
Accuracy	0.6741 95% CI [0.6536, 0.6942]
Kappa	0.333
Sensitivity	0.686
Specificity	0.667
PPV	0.536
NPV	0.792
Balanced Accuracy	0.677

SA3. Exclude subjects who endorsed a conflicting response variable elsewhere in the survey

A number of participants indicated elsewhere in the survey that they had also received outpatient mental health treatment in the last 12 months. It is possible that an individual had two episodes of mental illness within the 12 month period and that they obtained outpatient treatment for one of them and not the other. It is also possible that the inconsistency may reflect expected variation in attention when completing the survey; misinterpreting one (or both) of the questions; and/or human error when indicating survey responses. Therefore, we conducted a more conservative analysis in which we excluded participants that were missing all 9 DSM items and additionally excluded a further 6,606 participants who indicated elsewhere in the survey that they had also received outpatient mental health treatment in the last 12 months.

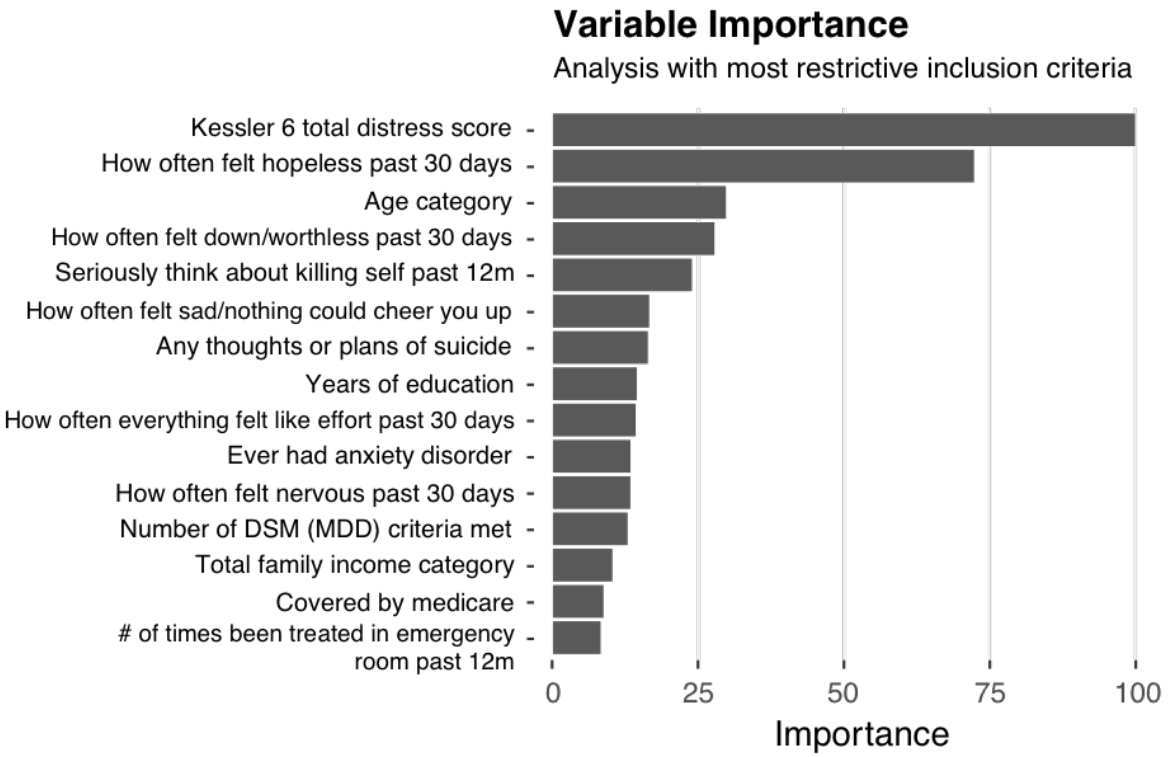
In this analysis, we included 5,669 participants (i.e. 12,291 – 6,606 (who said they received outpatient treatment in last year) – 16 (unknown or no response)). This included 4,677 participants in the 2008-2013 training set, and 992 participants in the 2014 external validation set. Amongst the training set, 37.9% of participants endorsed needing treatment but not getting it. In the validation set, 33.0% of participants endorsed needing treatment but not getting it. All statistical analyses were conducted as described in the main manuscript.

As for the previous sensitivity analysis, during cross-validation within the 2008-2013 data, model performance was above chance and broadly comparable to the performance metrics obtained in the main analysis. For example, balanced accuracy in this sample was 67.4% on average across folds and repeats (SD = 2.0%), compared to 70.5% obtained in the larger sample that included the subjects with missing data.

Supplementary Table 5: Cross-validated performance metrics for predicting engagement using maximally restrictive inclusion criteria

ST5	Mean	SD
AUC	0.761	0.019
Accuracy	0.705	0.019
Kappa	0.356	0.040
Sensitivity	0.547	0.029
Specificity	0.801	0.020
PPV	0.626	0.030
NPV	0.744	0.014
Balanced Accuracy	0.674	0.020

Supplementary Figure 3: Variable importance in predicting engagement amongst participants with most restrictive inclusion criteria



This model successfully generalized to predict engagement in the external validation set, again with performance comparable to the performance observed in the primary analysis. For instance, the balanced accuracy in the external validation sample here (69.5%) was approximately the same as for the main analysis (70.5%). As before, balanced accuracy was significantly better than chance, mean = 0.695, (exact binomial) 95% confidence interval [0.665, 0.723], $p<0.0001$. This suggests that even when we restrict the analysis to participants that were not missing DSM items and that were as reliable as possible in their responses, it was still possible to predict our outcome of interest in an external validation sample.

Supplementary Table 6: Performance metrics based on external validation of analysis with most restrictive inclusion criteria

ST6	
Accuracy	0.744 95% CI [0.7156, 0.7709]
Kappa	0.402
Sensitivity	0.550
Specificity	0.839
PPV	0.627
NPV	0.791
Balanced Accuracy	0.695

Supplementary Table 7. Permutation testing metrics for table 2 in the main manuscript.

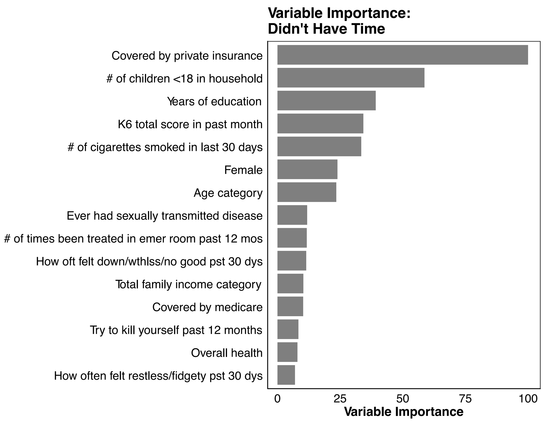
Supplementary Table 7. Permutation metrics for Table 2 in the main manuscript

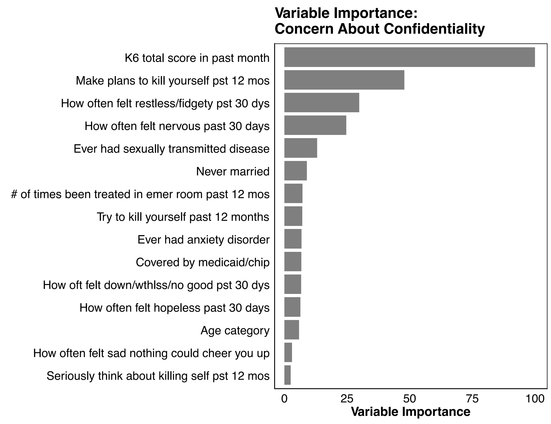
Reason	Endorsement Rates			Average Permuted (Null) Performance			True External Validation Performance			
	2008-2014	2008-2013	2014	BAC	Sens	PLR	BAC	Sens	PPV	PLR
Couldn't afford cost	47.7%	47.9%	46.4%	50% (2.1)	48% (2.2)	1.0 (0.09)	**64.2%**	**62.9%**	**61.2%**	**1.81**
Thought they could handle without treatment	22.2%	22.2%	22.4%	50% (1.8)	22% (2.8)	0.99 (0.16)	**55.8%**	**31.0%**	**31.5%**	**1.59**
Didn't know where to go for service	16.7%	16.0%	20.6%	50% (1.5)	16% (2.4)	0.97 (0.19)	**52.9%**	**20.6%**	**26.6%**	**1.40**
Some other reason	15.3%	15.0%	16.8%	50% (1.7)	15% (2.9)	1.0 (0.23)	51.8%	17.9%	20.1%	1.25
Thought might be committed or forced meds	15.2%	15.3%	14.8%	50% (2.6)	15% (4.4)	1.1 (0.35)	**64.9%**	**40.6%**	**39.5%**	**3.75**
Didn't have time/too busy	14.2%	14.3%	13.8%	50% (2.0)	14$ (3.5)	0.96 (0.28)	**56.2%**	**24.8%**	**24.1%**	**1.99**
Not enough health insurance coverage	11.7%	11.5%	13.1%	50% (1.6)	11% (2.8)	0.98 (0.29)	**55.3%**	**20.6%**	**23.6%**	**2.06**
Concerned about opinion of neighbors	11.0%	10.9%	11.8%	50% (1.9)	11% (3.3)	0.99 (0.35)	**56.3%**	**21.9%**	**24.0%**	**2.36**
Didn't think treatment would help	10.9%	10.9%	11.0%	50% (1.7)	11% (3.0)	0.98 (0.32)	**53.0%**	**16.1%**	**16.3%**	**1.58**
Concern about confidentiality	9.7%	9.7%	9.8%	50% (1.8)	10% (3.2)	1.1 (0.38)	**54.1%**	**16.8%**	**17.4%**	**1.93**
Don't think they needed it at that time	8.6%	8.6%	8.6%	50% (1.6)	8.7% (2.9)	1.0 (0.38)	**53.3%**	**14.5%**	**14.6%**	**1.82**
Concern about effect on job	8.1%	8.0%	8.4%	50% (1.6)	7.3% (3.0)	0.94 (0.42)	51.8%	11.1%	11.8%	1.47
Health insurance didn't cover it	6.5%	6.6%	6.1%	50% (1.7)	7.0% (3.3)	1.1 (0.56)	48.6%	3.4%	3.4%	0.55
Didn't want others to find out	6.5%	6.4%	6.8%	50% (1.7)	6.5% (3.2)	1.1 (0.56)	52.3%	10.6%	11.5%	1.77
Had no transportation or treatment too far	5.8%	5.6%	7.2%	50% (1.7)	6.1% (3.1)	1.1 (0.65)	52.5%	10.1%	13.2%	1.98

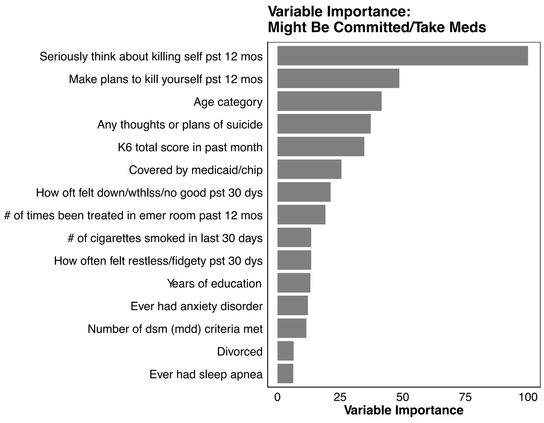
BAC = Balanced Accuracy. Sens = Sensitivity. PLR = Positive Likelihood Ratio. PPV = Positive Predictive Value. Permuted (Null) Performance columns reflect mean (sd) of performance metrics when the entire analytic pipeline was repeated 200 times using shuffled labels. Bold font indicates true external validation performance metrics that were significantly greater than 95% of all permuted metrics (i.e. 1-tailed alpha of 0.05).

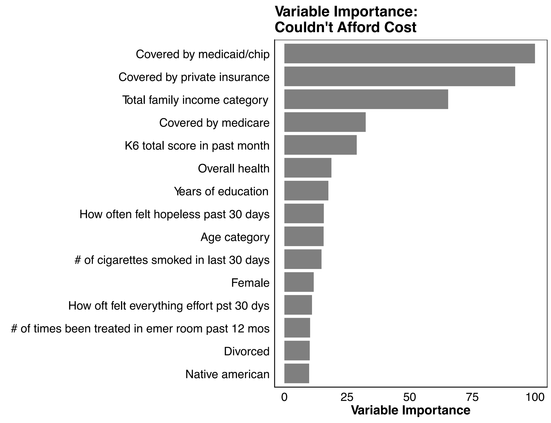
Supplementary Group Level Variable Importance Plots

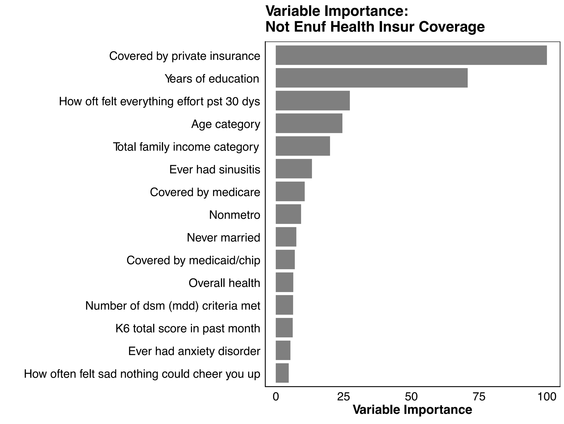
One figure for each of the 10 reasons that were predictable above chance.

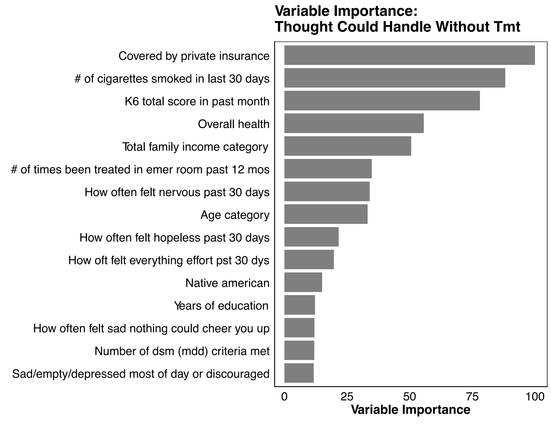


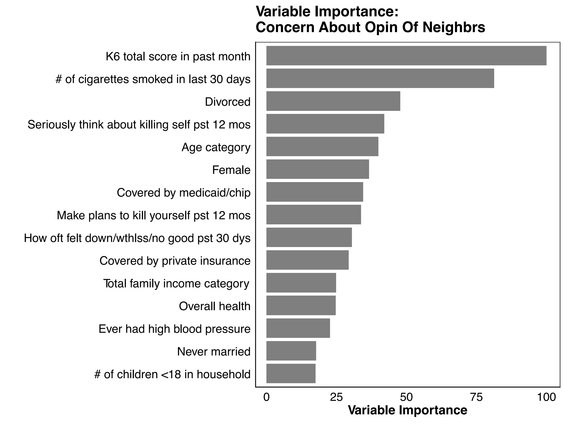


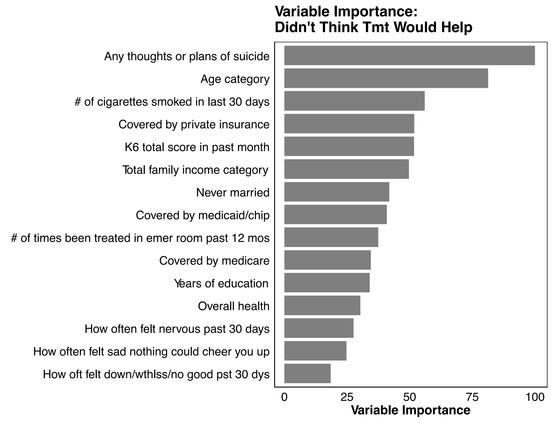


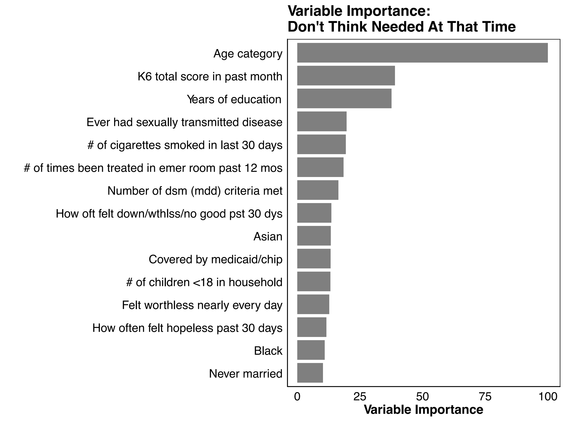


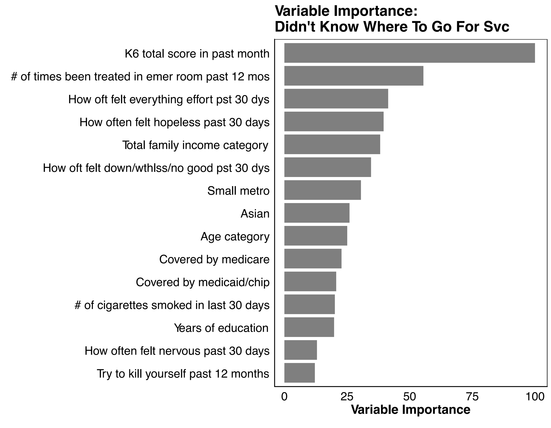












Supplementary Table 8. Alternative thresholds for identifying individuals at risk of not initiating treatment.

Different clinical contexts may impose different selection pressures concerning an acceptable rate of false positives vs false negatives. For example, if false positives are considered to be more costly, perhaps due to fatigue amongst providers receiving unnecessary notifications, then a higher threshold can be used to ensure that the model PPV is high enough to merit the notification. If a threshold of 0.7 is used then the model would have a PPV of 61%, rather than a PPV of 47% that is obtained when the less conservative threshold of 0.5 is selected.

In practice, it is highly likely that any implementation of this kind of model would in practice involve a custom selection of the threshold probability to control the number of times a prediction would be acted on (rather than FP/FN rates, for example). In other words, a system is likely to want to control the number of times that a clinical decision support notification is shown to care givers, or the number of times a care manager is told to intervene with a patient, on the basis of operational bandwidth i.e. how many interventions they can afford to deliver.

ST8. Alternative thresholds for identifying individuals at risk of not initiating treatment

Thresholds	Sensitivity	Specificity	PPV	NPV	BAC
0.05	99.80%	5.30%	29.30%	98.50%	52.60%
0.1	98.40%	14.30%	31.10%	95.70%	56.30%
0.15	97.00%	24.00%	33.40%	95.40%	60.50%
0.2	95.20%	31.80%	35.40%	94.40%	63.50%
0.25	92.80%	38.90%	37.40%	93.20%	65.90%
0.3	88.30%	44.80%	38.60%	90.70%	66.60%
0.35	86.20%	51.20%	41.00%	90.40%	68.70%
0.4	82.00%	56.40%	42.50%	88.80%	69.20%
0.45	78.30%	62.50%	45.00%	88.00%	70.40%
0.5	72.40%	68.50%	47.40%	86.30%	70.50%
0.55	66.50%	74.50%	50.60%	85.00%	70.50%
0.6	58.30%	79.90%	53.20%	83.00%	69.10%
0.65	50.20%	85.20%	57.10%	81.30%	67.70%
0.7	42.30%	89.50%	61.30%	79.80%	65.90%
0.75	32.80%	93.30%	65.70%	77.90%	63.00%
0.8	20.40%	96.10%	67.00%	75.40%	58.20%
0.85	8.30%	99.00%	75.70%	73.30%	53.60%
0.9	0.70%	99.90%	70.00%	71.90%	50.30%
0.931	0.10%	100.00%	100.00%	71.80%	50.10%

Supplementary Table 9. Breakdowns of the % of individuals who need treatment but do not receive it, by sex, age, and race. For example, 26.1% of men needed but failed to receive treatment, compared to 31.8% of women.

ST9. Breakdown of treatment initiation by sex, age, and race			
Participant characteristic		% that endorsed needing treatment but not receiving it	Number of individuals in group
All		30.2%	20785
Sex	Male	26.1%	5833
	Female	31.8%	14952
Age	18-25	36.6%	8638
	26-34	33.7%	3347
	35-49	27.4%	5254
	50-64	17.6%	2727
	65+	7.7%	819
Race	White	29.0%	15923
	Black/African American	34.5%	1364
	Native American	33.6%	301
	Native Hawaiian/Pacific Islander	19.5%	41
	Asian	31.6%	275
	Multi-racial	37.8%	783
	Hispanic	32.7%	2098

Supplementary Table 10. Breakdowns of reasons endorsed for not getting needed treatment, by sex, age, and race. All numbers are percentages. For example, 44.9% of men endorsed cost as a reason why they did not get treatment that they need, compared to 48.6% of women.

ST10. Breakdown of reasons endorsed for not getting needed treatment by age, race, and gender.															
	All	Sex		Age category					Race						
Reason		Male	Female	18-25	26-34	35-49	50-64	65+	White	Black	Native American	Native HI/Pac	Asian	Multi-racial	Hispanic
Couldn't afford cost	47.7	44.9	48.6	44.9	51.9	50.7	50.4	19.7	49.6	37.4	35.4	50.0	44.8	43.9	45.4
Thought they could handle without treatment	22.2	21.5	22.5	24.8	20.0	19.8	17.8	21.3	22.8	18.6	26.3	12.5	33.3	24.1	18.3
Didn't know where to go for service	16.7	16.8	16.7	18.3	15.7	15.3	14.2	8.2	16.2	16.5	15.2	12.5	25.3	19.4	18.8
Some other reason	15.3	16.5	14.9	14.0	14.4	15.7	23.0	26.2	15.4	13.9	19.2	0	19.5	19.4	12.6
Thought might be committed or forced to take meds	15.2	15.3	15.2	19.0	14.2	10.6	7.5	8.2	14.9	18.0	12.1	12.5	18.4	18.4	14.4
Didn't have time/too busy	14.2	10.5	15.4	14.9	16.4	13.5	7.9	6.6	14.6	12.1	19.2	12.5	13.8	16.0	11.9
Not enough health insurance coverage	11.7	10.0	12.2	9.0	12.6	15.2	17.2	11.5	12.4	8.7	5.1	12.5	18.4	9.9	10.0
Concerned about opinion of neighbors	11.0	14.0	10.1	13.3	11.4	8.1	5.4	3.3	10.8	10.8	11.1	37.5	12.6	13.9	11.0
Didn't think treatment would help	10.9	11.2	10.8	13.4	7.4	8.7	8.8	13.1	11.1	9.1	9.1	0	23.0	13.9	8.5
Concern about confidentiality	9.7	10.0	9.6	10.9	10.0	7.6	7.3	9.8	9.7	9.7	11.1	12.5	13.8	12.9	7.9
Don't think they needed it at that time	8.6	9.1	8.5	11.2	5.7	6.3	5.0	11.5	8.2	10.0	8.1	12.5	20.7	8.5	9.4
Concern about effect on job	8.1	9.2	7.7	7.2	10.4	9.6	4.8	0	7.9	9.1	6.1	25.0	9.2	7.8	8.6
Health insurance didn't cover it	6.5	6.6	6.4	6.5	7.2	5.6	8.2	1.6	6.2	8.2	2.0	12.5	11.5	7.8	6.7
Didn't want others to find out	6.5	6.1	6.6	8.2	5.9	4.3	3.8	0	6.3	7.1	7.1	0	16.1	9.2	4.6
Had no transportation or treatment too far	5.8	4.7	6.2	6.0	5.9	5.2	7.3	1.6	5.5	5.6	14.1	0	9.2	8.8	5.4

All numbers are percentages. For example, 44.9% of men endorsed cost as a reason why they did not get treatment that they need, compared to 48.6% of women.

R Package Citations

This manuscript relied heavily on user-written R packages, credited below.

High Performance Computing:

{doMC} Revolution Analytics (2014). doMC: Foreach parallel adaptor for the multicore package. R package version 1.3.3. http://CRAN.R-project.org/package=doMC

Plotting/visualisation tools and data manipulation:

{ggplot2} H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

{dplyr} Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. http://CRAN.R-project.org/package=dplyr

{daff} Paul Fitzpatrick and Edwin de Jonge (2016). daff: Diff, Patch and Merge for Data.frames. R package version 0.2.0. http://CRAN.R-project.org/package=daff

Machine Learning:

{caret} Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem and Luca Scrucca. (2015). caret: Classification and Regression Training. R package version 6.0-52. http://CRAN.R-project.org/package=caret

{glmnet} Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. http://www.jstatsoft.org/v33/i01/

Complete variable list

Short-form (as in NSDUH)	***Meaning***
adult	Aged over 18?
CATAG6	Age category recoded (6 levels)
sex	1= male, 2=female
IRFAMIN3	Recoded - imputation revised - total family income
COUTYP2	County type
NEWRACE2	Race/hispanicity recode (7 levels)
NRCH17_2	Recoded # of children < 18 in household
IRMARIT	Imputation revised marital status
incomeHiLo	household income over 40k per year?
MEDICARE	Covered by medicare?
CAIDCHIP	Covered by medicaid/chip?
PRVHLTIN	Covered by private insurance?
IREDUC2	Recoded - imputation revised education
AD_MDEsum	Total number of AD_MDE A1:A9 items endorsed
AD_MDEA1	Sad/empty/depressed most of day or discouraged
AD_MDEA2	Lost interest or pleasure in most things
AD_MDEA3	Changes in appetite or wt
AD_MDEA4	Sleep problems
AD_MDEA5	Others noticed that r was restless or lethargic
AD_MDEA6	Felt tired/low energy nearly every day
AD_MDEA7	Felt worthless nearly every day
AD_MDEA8	Inability to concentrate or make decisions
AD_MDEA9	Any thoughts or plans of suicide
SUICTHNK	Seriously think about killing self pst 12 mos
SUICPLAN	Make plans to kill yourself pst 12 mos
SUICTRY	Try to kill yourself past 12 months
K6SCMON	Kessler 6 total score in past month
DSTCHR30	How often felt sad nothing could cheer you up
DSTEFF30	How oft felt everything effort pst 30 dys
DSTHOP30	How often felt hopeless past 30 days
DSTNGD30	How oft felt down/wthlss/no good pst 30 dys
DSTNRV30	How often felt nervous past 30 days
DSTRST30	How often felt restless/fidgety pst 30 dys
cig30	Number of days in the last 30 days that you smoked cigarettes
NMERTMT2	# of times been treated in emergency room past 12 mos
LIFANXD	Ever had anxiety disorder

LIFASMA	Ever had asthma
LIFBRONC	Ever had bronchitis
LIFCIRR	Ever had cirrhosis of the liver
LIFDIAB	Ever had diabetes
LIFHARTD	Ever had heart disease
LIFHEPAT	Ever had hepatitis
LIFHBP	Ever had high blood pressure
LIFPNEU	Ever had pneumonia
LIFSTDS	Ever had sexually transmitted disease
LIFSINUS	Ever had sinusitis
LIFSLPAP	Ever had sleep apnea
LIFSTROK	Ever had stroke
LIFULCER	Ever had ulcer or ulcers
HEALTH	Overall self-reported health status (1-5 likert)
DEPRSYR	Had depression in the last 12 months
AUOPTYR	Rcvd outpatient mh trmt pst 12 mos
AUUNMTYR	Needed mh trmt but didn't get it past 12 mos
AUUNCOST	No mh tmt couldn't afford cost
AUUNNBR	No mh tmt concern about opin of neighbrs
AUUNJOB	No mh tmt concern about effect on job
AUUNNCOV	No mh tmt health insur didn't cover
AUUNENUF	No mh tmt not enuf health insur coverage
AUUNWHER	No mh tmt didn't know where to go for svc
AUUNCFID	No mh tmt concern about confidentiality
AUUNCMIT	No mh tmt might be committed/take meds
AUUNNOND	No mh tmt don't think needed at that time
AUUNHNDL	No mh tmt thought could handle without tmt
AUUNNHLP	No mh tmt didn't think tmt would help
AUUNBUSY	No mh tmt didn't have time
AUUNFOUT	No mh tmt didn't want others to find out
AUUNNTSP	No mh tmt had no transportation or tmt too far
AUUNSOR	No mh tmt for some other reason

References

1 Kazdin AE. Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behav Res Ther* 2017; **88**: 7–18.

2 Tang TZ, DeRubeis RJ. Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *J Consult Clin Psychol* 1999; **67**: 894–904.

3 Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. *Bull World Health Organ* 2004; **82**: 858–66.

4 Friedman JH. Stochastic Gradient Boosting. *Comput Stat Data Anal* 1999; **38**: 367–78.

5 Ridgeway G. Generalized Boosted Models : A guide to the gbm package. *Compute* 2007; **1**: 1–12.

6 Chekroud AM, Zotti RJ, Shehzad Z, *et al.* Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 2016; **3**: 243–50.

7 Koutsouleris N, Kahn RS, Chekroud AM, *et al.* Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry* 2016; **3**: 935–46.