Cross-trial prediction of treatment outcome in depression: a machine learning approach

Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, Philip Robert Corlett

Summary

Background Antidepressant treatment efficacy is low, but might be improved by matching patients to interventions. At present, clinicians have no empirically validated mechanisms to assess whether a patient with depression will respond to a specific antidepressant. We aimed to develop an algorithm to assess whether patients will achieve symptomatic remission from a 12-week course of citalopram.

Methods We used patient-reported data from patients with depression (n=4041, with 1949 completers) from level 1 of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D; ClinicalTrials.gov, number NCT00021528) to identify variables that were most predictive of treatment outcome, and used these variables to train a machine-learning model to predict clinical remission. We externally validated the model in the escitalopram treatment group (n=151) of an independent clinical trial (Combining Medications to Enhance Depression Outcomes [COMED]; ClinicalTrials.gov, number NCT00590863).

Findings We identified 25 variables that were most predictive of treatment outcome from 164 patient-reportable variables, and used these to train the model. The model was internally cross-validated, and predicted outcomes in the STAR*D cohort with accuracy significantly above chance (64·6% [SD 3·2]; p<0·0001). The model was externally validated in the escitalopram treatment group (N=151) of COMED (accuracy 59·6%, p=0.043). The model also performed significantly above chance in a combined escitalopram-buproprion treatment group in COMED (n=134; accuracy 59·7%, p=0·023), but not in a combined venlafaxine-mirtazapine group (n=140; accuracy 51·4%, p=0·53), suggesting specificity of the model to underlying mechanisms.

Interpretation Building statistical models by mining existing clinical trial data can enable prospective identification of patients who are likely to respond to a specific antidepressant.

Funding Yale University.

Lancet Psychiatry 2016; 3: 243–50

Published Online
January 20, 2016
http://dx.doi.org/10.1016/S2215-0366(15)00471-X

See Comment page 192

Department of Psychology (A M Chekroud MSc, Z Shehzad MSc, M K Johnson PhD, T D Cannon PhD), **Department of Biostatistics** (R Gueorguieva PhD), **and Department of Psychiatry** (T D Cannon, J H Krystal MD, P R Corlett PhD), **Yale University, New Haven, CT, USA; Capital One, McLean, VA, USA** (R J Zotti BSc); **Department of Psychiatry, UT Southwestern, Dallas, TX, USA** (M H Trivedi MD); **and Centre for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA** (A M Chekroud)

Correspondence to:
Mr Adam M Chekroud,
Department of Psychology,
Yale University, New Haven,
CT 06511, USA
adam.chekroud@yale.edu

Introduction

As few as 11–30% of patients with depression reach remission with initial treatment, even after 8–12 months.[1–4] One factor reducing effectiveness of treatment is the inability to personalise pharmacotherapy.[5] Clinicians match patients with specific antidepressants via a prolonged period of trial and error, delaying clinical improvement and increasing risks and costs of treatment. The absence of clinical prediction tools in psychiatry starkly contrasts with other areas of medicine, such as oncology, cardiology, and critical care, where algorithmic models often have important roles in medical decision making,[6–8] and routinely outperform judgment of individual clinicians.[9–11]

A challenge when developing predictive clinical tools is to establish what information should be used. Genetic and brain imaging measures are possible sources of information, and have generated interest.[12,13] However, even if effective, the cost and time of collecting and processing data might not be practical. By contrast, behavioural (eg, patient-reported) data are already collected but perhaps underused. Clinical experience guides what information is used in treatment decisions;[14] however, early-stage clinicians have little experience, and even experienced clinicians might overlook useful information or overweight salient clinical examples.[15] Previous attempts to identify clinical predictors of treatment outcome have generally identified a few predictors based on clinical experience, and have investigated their overall effect in a stepwise manner.[16] One important study took a slightly different approach, quantifying the effect of nine symptom dimensions derived from a factor analysis.[17] However, examination of all potential predictors simultaneously in an unbiased manner (sometimes called data mining) provides an opportunity for discovery. Machine-learning methods are especially well suited for this challenge.[18,19] Rather than separately considering the effect of one variable on an outcome of interest, machine-learning methods identify patterns of information in data that are useful to predict outcomes at the individual patient level. Modern machine-learning approaches offer key benefits over traditional statistical approaches (generalised linear models, and even non-linear regression models [generalised additive models]), because (when present) they can detect complex (non-linear) high-dimensional interactions that might inform predictions.

Research in context

Evidence before this study

In major depressive disorder, prediction of treatment outcome is an important goal because most patients do not reach remission with their first course of treatment. We searched PubMed from inception to Aug 6, 2015 with the terms (“depression” OR “major depressive disorder”) AND “prediction” AND “outcome” in any field, with no language restrictions. We retrieved and scanned 734 articles, then focused on the 225 articles in which (“depression” OR “major depressive disorder”) was in the title. All articles that we deemed not to be relevant on the basis of their titles were excluded. Abstracts of the remaining articles were reviewed to identify potentially relevant articles, and, on the basis of this selection, we read full-text articles.

Typically, researchers examine the effect of a small number (eg, <30) of preselected predictor variables. By preselecting variables, novel predictive associations can be overlooked. Although 11 studies included sample sizes larger than 500, a crucial challenge for the specialty is to show that predictive models are accurate outside their discovery context. We identified only one study that developed a model in a large clinical trial sample and directly examined the validity of their model in a large independent sample.

Added value of this study

Our study offers a data-driven method to identify useful predictor variables among a large number of candidate predictors, and to combine them for individual-patient predictions. We describe a machine-learning model optimised to detect future responders for a specific, first-line antidepressant (citalopram), with a simple 10-min questionnaire. The model’s accuracy is significantly above chance in a large clinical trial cohort; externally validated in a large, independent clinical trial cohort; and compares favourably to the accuracy of a pilot sample of psychiatrists. The model uses easy-to-obtain (patient-reportable) information, and could be hosted online or in the clinic using a mobile device, laptop, or desktop computer.

Implications of all the available evidence

We show that machine learning techniques applied to self-report questionnaire data can aid prediction of clinical remission for a specific antidepressant. This approach can easily be extended to include other sources of data (ie, biomarkers) for prediction—which might improve performance—and other treatments and clinical populations. Mining existing clinical trial data with these methods should enable patients to be matched with specific drug treatments, and these models warrant further investigation in prospective controlled trials.

We used a machine-learning approach to predict whether a patient will reach clinical remission from a major depressive episode with a 12-week course of citalopram.

Methods

Study design and clinical trial data

With data from a large, multicentre clinical trial of major depressive disorder (STAR*D), we built a predictive model, and internally cross-validated the model. We externally validated the model developed in STAR*D in a wholly independent clinical trial cohort consisting of three independent treatment groups (COMED).

The STAR*D trial (ClinicalTrials.gov, number NCT00021528) is the largest prospective, randomised controlled study of outpatients with major depressive disorder. Patients were recruited from primary and psychiatric care settings in the USA from June, 2001, to April, 2004. Study protocols have been described in detail previously.[1,16,20] Eligible participants were treatment-seeking outpatients, with a primary clinical (DSM-IV) diagnosis of non-psychotic major depressive disorder, a score of at least 14 on the 17-item Hamilton Depression Rating Scale (HAM-D), and aged 18–75 years.[16] Since our study sought to predict initial antidepressant response, we focused on the first treatment stage—a 12-week course of citalopram, a commonly used SSRI antidepressant.

See Online for appendix

The COMED trial (ClinicalTrials.gov, number NCT00590863) was a single-blind, randomised, placebo-controlled trial comparing efficacy of medication combinations in the treatment of major depressive disorder. Briefly, 665 outpatients were enrolled between March, 2008, and February, 2009, across six primary care sites and nine psychiatric care sites.[21] Eligible patients were aged 18–75 years, had a primary DSM-IV-based diagnosis of non-psychotic major depressive disorder, had recurrent or chronic depression (current episode ≥2 years), and had a score of at least 16 on the 17-item HAM-D rating scale. Exclusion criteria included all patients who had comorbid psychotic illness or bipolar disorder, or who needed admission to hospital. Patients were randomly allocated (1:1:1) to one of the following three groups: escitalopram plus placebo (monotherapy); escitalopram plus buproprion; or venlafaxine plus mirtazapine.

Data for both studies were acquired from the US National Institute of Mental Health (NIMH) through limited access data use certificates, as detailed in the appendix (p 1). The appendix contains a detailed description of the statistical modelling pipeline and full inclusion and exclusion criteria for both trials.

Dataset description

We took a complete cases approach, including only patients without missing observations. Although patients in both trials were encouraged to visit the clinic every 2 weeks, most patients did not attend every appointment. Our analyses focused on patients for whom a severity score was recorded after 12 or more weeks of treatment. Of the original 4041 patients in STAR*D, 12 had no outcome data, and 2044 completed fewer than 12 weeks of treatment, leaving 1985 patients. For COMED, of the original 665 patients, two had no outcome data, and 187 did not complete 12 weeks, leaving 476 patients. We excluded patients with missing baseline data (36 patients in STAR*D and 51 patients in COMED). We trained the model using these final 1949 patients from STAR*D. 425 patients were included for external validation in COMED (escitalopram-placebo 151; buproprion-escitalopram 134; venlafaxine-mirtazapine 140). Baseline depressive severity was similar for the final completer sample (mean Quick Inventory of Depressive Symptomatology [QIDS] severity 15·1, range 2–27, IQR 12–18) and those excluded for missing outcome data (mean 15·9, range 2–27, IQR 13–19). Supplementary last observation carried forward analyses of the STAR*D dataset (n=3518) are documented in the appendix.

Clinical outcomes

We assessed outcome measurements according to the 16-item self-report QIDS (QIDS-SR$_{16}$). We focused on the key clinical target of clinical remission (final score ≤5 in the 16-item QIDS-SR$_{16}$), in line with previous studies,[1,16] since it is associated with better function and a better prognosis than response without remission. Since visit weeks differed between the two trials, for STAR*D, final scores were the last available measurement at either 12 or 14 weeks; for COMED, final scores were at 12 or 16 weeks. Prediction of other clinical outcomes of interest (eg, percentage symptom change) would be possible with the same methods.

Model development

We constructed and examined all models with repeated ten-fold cross-validation (ten repeats), which partitions the original sample into ten disjoint subsets, uses nine of those subsets in the training process, and then makes predictions about the remaining subset. To avoid opportune data splits, we averaged model performance metrics across test folds. For external validation, we applied the final model built in the STAR*D cohort without modification to predict treatment outcomes in each COMED treatment group separately.

Predictor selection

We extracted all readily available sources of information that overlapped for patients in both STAR*D and COMED trials. Information included various sociodemographic features, DSM-IV-based diagnostic items, depressive severity checklists (eg, QIDS-SR and HAM-D), eating disorder diagnoses, whether the patient had previously taken specific antidepressant drugs, the number and age of onset of previous major depressive episodes, and the first 100 items of the psychiatric diagnostic symptom questionnaire.[22] We included 164 variables. Unfortunately, several additional items in the STAR*D dataset were not available in COMED, and vice versa, so they could not be used. The full list of overlapping variables in the two trials is detailed in the appendix.

A key challenge for prediction is to identify which variables to use. A classic solution to this problem is to use a stepwise feature selection procedure,[23] but this approach is slow and prone to over-fitting.[24,25] We assessed all 164 predictors simultaneously with a

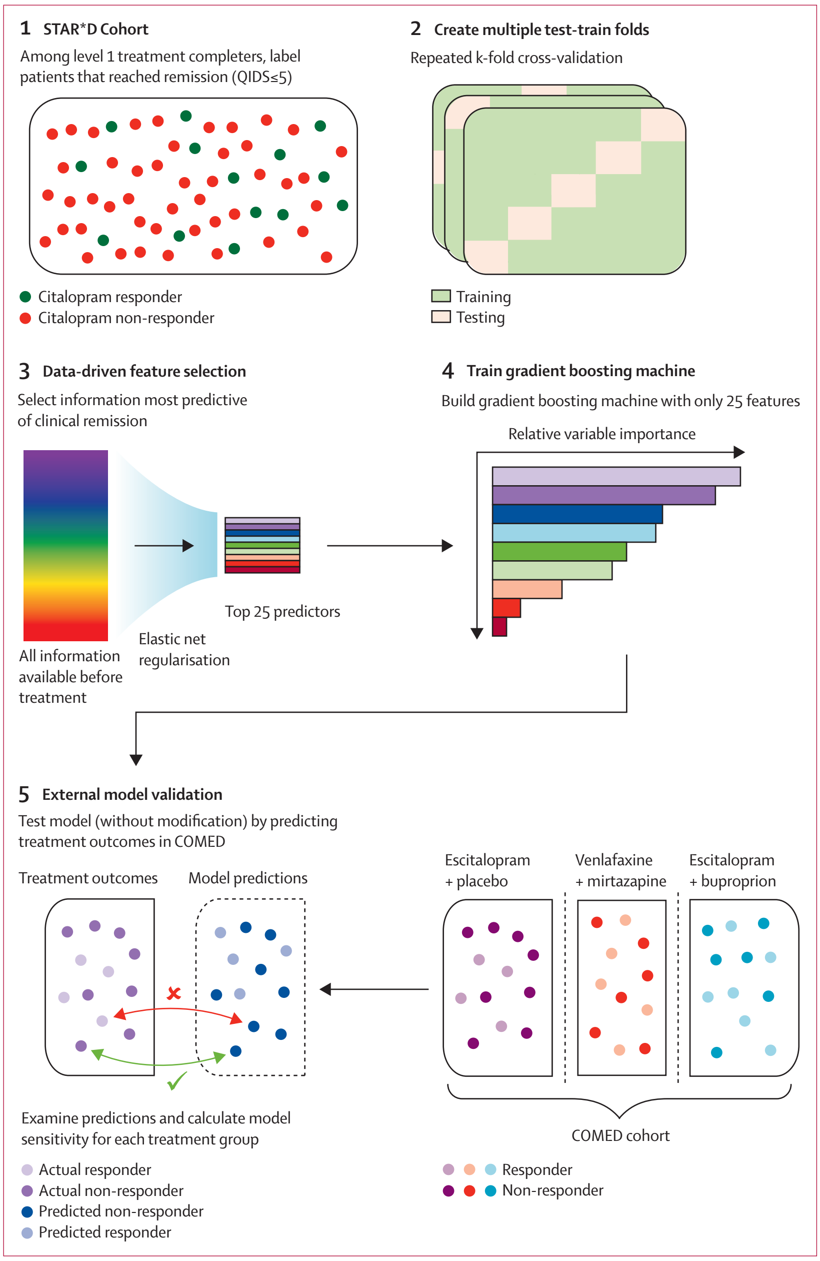
1 STAR*D Cohort
Among level 1 treatment completers, label patients that reached remission (QIDS≤5)
● Citalopram responder
● Citalopram non-responder
2 Create multiple test-train folds
Repeated k-fold cross-validation
Training
Testing
3 Data-driven feature selection
Select information most predictive of clinical remission
All information available before treatment
Elastic net regularisation
Top 25 predictors
4 Train gradient boosting machine
Build gradient boosting machine with only 25 features
Relative variable importance
5 External model validation
Test model (without modification) by predicting treatment outcomes in COMED
Treatment outcomes
Model predictions
Escitalopram + placebo
Venlafaxine + mirtazapine
Escitalopram + buproprion
Examine predictions and calculate model sensitivity for each treatment group
COMED cohort
● Actual responder
● Actual non-responder
● Predicted non-responder
● Predicted responder
● ● ● Responder
● ● ● Non-responder

Figure 1: **Analysis pipeline**

In the STAR*D cohort, label level 1 treatment completers according to whether they reached remission or not (1). Set up ten repeats of ten-fold cross-validation (2). Identify predictors that are most predictive of treatment outcome with a data-driven selection method (elastic net regularisation; 3). Use top 25 predictive features to train a machine-learning algorithm to predict treatment outcomes for citalopram (4). Examine model performance in three treatment groups of an independent clinical trial cohort (COMED; 5). QIDS=quick inventory of depressive symptomatology.

	Coefficient
Initial QIDS total severity	0·07793
Currently employed	–0·06946
QIDS psychomotor agitation	0·06929
QIDS energy or fatiguability	0·05893
Black or African American	0·05559
Initial HAM-D depressive severity	0·05290
QIDS mood (sad)	0·04895
Years of education	–0·04712
HAM-D loss of insight	–0·04625
HAM-D somatic energy	0·03658
HAM-D somatic anxiety	0·03312
Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?	0·03034
HAM-D delayed insomnia	0·02992
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	0·02673
Did you try to avoid activities, places, or people that reminded you of a traumatic event?	0·02651
White	–0·02593
Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Standing in long lines	0·02477
Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Driving or riding in a car	0·02424
Have you been bothered by aches and pains in many different parts of your body?	0·02249
HAM-D suicide	0·02175
Depressed mood most of the day, nearly every day	0·02095
Did you have attacks of anxiety that caused you to avoid certain situations or to change your behaviour or normal routine?	0·01989
Ever taken sertraline	0·01851
Number of previous major depressive episodes	0·01832
QIDS sleep onset insomnia	0·01819

QIDS=Quick Inventory of Depressive Symptomatology. HAM-D=Hamilton Depression Rating Scale.

Table 1: **Top 25 predictive items in elastic net model**

method that avoids issues of correlated predictors and over-fitting (elastic net regularisation).[26,27] The method has two primary effects: coefficients of correlated predictors are shrunk towards each other, and uninformative features are removed from the model. We used the elastic net model to select the 25 best features from those available using the STAR*D training sample. The concept of nuisance covariates does not apply since all information extracted from the trial was included in the model (that is, all information was of interest). This two-step procedure of preselecting variables before final model building enabled us to ensure that the final predictive model would need only 25 variables, a number that was selected to balance practical usability with model performance. This approach should only be used when an independent validation set is available, otherwise it may introduce bias into error estimates.[19] We also assessed smaller models, using 10 or 15 predictors (appendix).

Predictive model building

We used the 25 predictive features to train a machine-learning algorithm to predict clinical remission. We used a gradient boosting machine, from a class of powerful machine-learning approaches showing success in a range of applications.[28–30] Rather than fitting one strong model to a dataset, a gradient boosting machine is built by combining several weakly predictive models to relate the predictors and outcome.[31] Crucially, when each successive model is fit, the model focuses on the data that previous models failed to predict. We fitted a tree-based ensemble to the top 25 predictors identified by the elastic net.

We developed a model to detect patients for whom citalopram is beneficial (rather than predicting non-responders). We selected optimum tuning parameters during cross-validation through an area under the receiver-operating curve (ROC)-maximisation process (comparing true positives to false positives). We used the best performing model in the training dataset to generate predictions in the independent validation set. We measured the significance of the model's accuracy with a one-tailed binomial test of model accuracy relative to the bigger class proportion (null-information rate).[32,33] We also measured other relevant descriptions of model discrimination—including sensitivity, specificity, and area under curve (AUC)—at each stage. Figure 1 illustrates the analysis pipeline. All analyses were implemented in R (version 3.1.2). All R-code we developed for statistical modelling is available upon request.

For more on **R-code** see http://cran.r-project.org

Role of the funding source

No funding source had any role in the study design, data collection, data analysis, data interpretation, writing, or submission of this report. The corresponding author had full access to all the data in the study. All authors had the final responsibility for the decision to submit for publication.

Results

We selected 25 predictors of remission or non-remission according to ranked absolute beta weights in the elastic net model (table 1). The top three predictors of non-remission were baseline QIDS-SR depression severity, feeling restless during the past 7 days (QIDS-SR psychomotor agitation), and reduced energy level during the past 7 days (QIDS-SR energy and fatiguability). The top three predictors of remission were currently being employed, total years of education, and loss of insight into one's depressive condition (HAM-D loss of insight). Although it is reasonable to interpret the relative magnitude of predictor coefficients in this case, it is difficult to interpret their direction in this highly multivariate, penalised regression model.

We built a machine-learning model with this restricted set of 25 variables. Table 2 contains performance measures of the model during internal cross-validation. The model achieved an average AUC of 0·700 (SD 0·036), suggesting sufficient predictive signal in the 25 questions selected by the elastic net. The majority class was non-remission, comprising 51·3% of patients (null information rate). Overall, the model's predictions had significant accuracy in predicting outcome in STAR*D patients (accuracy 64·6% [SD 3·2]; $p<9{\cdot}8\times10^{-33}$). The model prospectively identified 62·8% (SD 5·1) of patients who eventually reached remission (ie, sensitivity), and 66·2% (SD 4·6) of non-remitters (ie, specificity). Correspondingly, the model had a positive predictive value (PPV) of 64·0% (SD 3·5), and a negative predictive value (NPV) of 65·3% (SD 3·3). Model calibration and ROC curves are provided in the appendix. Results for smaller models, with only ten or 15 predictors, are discussed in the appendix.

To confirm the model's external generalisability, we applied the 25-item model from the STAR*D citalopram completers (without modification) to patients in three COMED treatment groups separately. The pattern of cross-trial model performance is shown in figure 2, and full model performance metrics are provided in the appendix. The model showed significant predictive performance in both the escitalopram-placebo group (79 remissions; accuracy 59·6%, 95% CI 51·3–67·5, p=0·043; PPV 65·0%, NPV 56·0%), and escitalopram-buproprion group (66 remissions; accuracy 59·7%, 50·9–68·1, p=0·023; PPV 59·7%, NPV 59·7%), but not the venlafaxine-mirtazapine group (72 remissions; accuracy 51·4%, 42·8–60·0; p=0·53; PPV 53·9%, NPV 50·0%). These differences in the predictability of treatment response occur even though the overall treatment efficacies were similar in the three groups (escitalopram-placebo 52·3%, escitalopram-buproprion 49·3%, venlafaxine-mirtazapine 51·4%). For completeness, we applied the same general modelling pipeline to each COMED treatment group separately (appendix). Although both the smaller ten-item and 15-item models performed well in the STAR*D cohort (AUC>0·683), neither showed significant performance in the escitalopram group of COMED (p>0·06; appendix).

We used a parallel approach to predict STAR*D patients' final QIDS-SR scores directly, rather than by use of remission versus non-remission status. We identified substantial overlap in the top 25 variables identified: the regression and classification models share 16 of their top 25 predictors, and eight of the top ten remain the same. In this format, our STAR*D model explained 17·5% of the variance in final QIDS-SR scores (root mean square error [RMSE] 4·54 [SD 0·20], R^2=0·175 [SD 0·052]). Including as one of the 25 variables the total QIDS-SR score measured at 2 weeks after baseline substantially improved all cross-validated performance measures in STAR*D completers relative to the baseline-only model (appendix; classification mean performance: accuracy 67·9% [SD 3·8], ROC 0·743 [0·041], sensitivity 69·1% [5·2]; regression equivalent: RMSE 4·17 [0·23], R^2=0·256 [0·060]).

	STAR*D (citalopram; internal cross-validation)	COMED (external validation)		
		Escitalopram plus placebo	Escitalopram plus buproprion	Venlafaxine plus mirtazapine
Accuracy	64·6% (3·2)	59·6%	59·7%	51·4%
AUC	0·700 (0·036)	..	..	..
p value (accuracy>NIR)	$<9{\cdot}8\times10^{-33}$	0·043	0·023	0·53
Sensitivity	62·8% (5·1)	49·4%	56·1%	38·9%
Specificity	66·2% (4·6)	70·8%	63·2%	64·7%
PPV	64·0% (3·5)	65·0%	59·7%	53·9%
NPV	65·3% (3·3)	56·0%	59·7%	50·0%

Data are mean (SD). AUC=area under receiver operating characteristic curve. NIR=null information rate. PPV=positive predictive value. NPV=negative predictive value.

Table 2: **Model performance during training and validation**

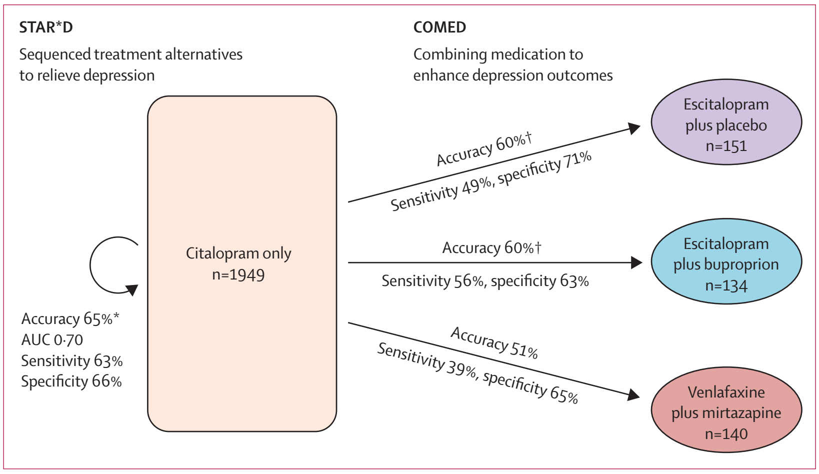
STAR*D
Sequenced treatment alternatives to relieve depression
COMED
Combining medication to enhance depression outcomes
Citalopram only
n=1949
Accuracy 65%*
AUC 0·70
Sensitivity 63%
Specificity 66%
Accuracy 60%†
Sensitivity 49%, specificity 71%
Accuracy 60%†
Sensitivity 56%, specificity 63%
Accuracy 51%
Sensitivity 39%, specificity 65%
Escitalopram plus placebo
n=151
Escitalopram plus buproprion
n=134
Venlafaxine plus mirtazapine
n=140

Figure 2: **Cross-trial prediction of antidepressant treatment outcomes**

Arrows indicate where a model was trained (arrow origin), and tested (arrow head). *p<0·001. †p<0·05.

Discussion

We developed a model to predict symptomatic remission after taking citalopram, a common antidepressant, with clinical rating data. Our model performance is similar to that of calculators of disease risk, recurrence, or treatment response in various areas of medicine, including oncology and cardiovascular disease.[34–37] In the context of depression, the model performs comparably to the best available biomarker—an EEG-based index[38,39]—but is less expensive, easier to implement, and validated in large internal and external clinical trial samples (a direct comparison is not possible owing to the different patient samples). The model was optimised to detect future responders, and improved substantially if a self-reported measure of overall depressive severity after 2 weeks of treatment was included in the model, indicating the possible usefulness of a 2-week prediction update (appendix).

A personalised medicine approach to pharmacotherapy holds promise in treatment of depression, a highly heterogeneous illness[40] for which no single treatment is universally effective, and for which many patients undergo several treatments before an appropriate regimen is identified. From large-scale clinical trials (including STAR*D and COMED), at a population level, about 30% of patients achieve symptomatic remission for a given treatment and episode.[1,21] However, personalised medicine shifts focus away from population remission rates and general drug efficacy, and tries to identify which 30% of patients are the best candidates for a specific drug. As an example, although remission rates for all drug treatments were similar (48–52%), our ability to prospectively predict treatment outcome was not (51–65%). Of course, development of generally effective and rapidly effective antidepressant treatments would be important advances for public health. Until then, development and implementation of innovative statistical methods to choose the best available drug for each patient offers an interim solution.[5,41,42] These findings are a step in the direction of precision medicine for psychiatry, but performance remains modest compared with that in other areas of medicine.

The success of these models depends on their ability to generalise. We took two important precautions. First, all variable selection and model building occurred during a repeated ten-fold cross-validation procedure. Second, we examined how the model trained on a large citalopram cohort would perform in other clinical trial cohorts with other treatment protocols, with differing recruitment criteria and distributions of symptoms. The external validation analysis showed that a citalopram model trained in the STAR*D cohort accurately predicted outcomes for the escitalopram treatment group of COMED. The model also showed significant accuracy in the escitalopram-buproprion group, but not in a combination SNRI group (venlafaxine-mirtazapine). This result shows that the model can successfully generalise to a completely independent sample, and has some degree of treatment specificity. The fact that the model failed to accurately predict response to the venlafaxine-mirtazapine group suggests that the model was not simply predicting a generic treatment response profile, nor was it predicting regression to the mean (or performance would have equivalent for all three COMED groups). Our use of wholly independent validation cohorts also showed that, although fewer predictors might still confer comparable model performance in the STAR*D cohort, these smaller models did not generalise to the escitalopram group of an independent clinical trial (appendix), highlighting the importance of external validation.[17]

At a minimum, statistical (and biomarker) models must show above chance performance to be useful. In our STAR*D analyses, an accuracy of 53·13% would have statistically outperformed the chance accuracy of 51·3% with this sample size, by conventional standards ($p<0·05$). Our model achieved an accuracy of 64·6%, surpassing this benchmark substantially. By contrast, clinical prediction of who will respond to which treatment is typically poor.[38] Similarly, in a pilot sample of psychiatrists and residents, the mean accuracy of 23 clinicians in predicting treatment outcome for 26 STAR*D patients was 49·3%, (where chance was 53·9%; appendix).

Our study has some limitations. Our assessment of clinician's ability to predict outcomes was preliminary, and future work will be needed to better quantify the accuracy of clinical judgment. At present, our model is not sufficiently powerful to justify withholding medication from a predicted non-responder, especially since medication could feasibly be the optimum treatment option for a predicted non-responder. However, a positive prediction is not trivial: compared with the baseline rate of antidepressant response, it more than doubles our confidence that a given patient is going to respond to citalopram (which could potentially confer additional placebo benefits).[43] Performance could be improved by including a greater selection of clinical or behavioural variables—we did not have some demographic variables (eg, income) that have previously been associated with treatment outcome in univariate analyses[16]—and perhaps further still with genetic or brain-based measures together with patient-reported data when training machine-learning algorithms. Crucially, this study offers a statistical pipeline to identify useful predictors from a large number of variables and combine them for clinical prediction.

Firm conclusions cannot be made about the model's ability to predict differential responses to various drugs studied. No pure placebo condition was used in these trials, and sample sizes in each COMED treatment group are small relative to STAR*D. Future work should assess the extent to which the differential predictive power of this approach for distinct medications reflects the distinctive neural mechanisms probed in these clinical trials, heterogeneity in the underlying neurobiology of depression among the patients who entered the trials, or indeed a more general detection of non-response.

More broadly, the ultimate goal is to identify choice-markers: that is, markers that simultaneously predict response to drug A, and non-response to drug B. Until then, we are guiding choice by sequentially identifying responders and non-responders for a specific drug (or drug combination). With a large enough dataset, including several treatment options (including non-pharmacological interventions),[43,44] such a methodological approach should be useful to develop an algorithm that matches patients to the best treatment option among alternatives. In the immediate term, a machine-learned model offers clinicians a quick and accessible tool to predict whether a specific patient will respond to citalopram. The success of this general approach depends crucially on collection and sharing of large-scale, clinical-grade datasets.

Contributors

AMC had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. AMC obtained funding. AMC, JHK, and PRC devised the study concept and design. AMC, PRC, JHK, and MKJ drafted the manuscript. AMC, PRC, JHK, TDC, RG, MKJ, and MHT critically revised the manuscript for important intellectual content. AMC, RJZ, RG, and ZS did the statistical analysis. All authors acquired, analysed, and interpreted data, and provided administrative, technical, and material support. PRC, JHK, TDC, and RG supervised the study.

Declaration of interests

RJZ is a data scientist at Capital One. He contributed in his personal capacity, and this manuscript does not reflect the views of his employer. RJZ holds stock in Capital One. MHT has served as an adviser or consultant to Abbott, AbdiIbrahim, 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 Pharmaceutica 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, NARSAD, NIMH, NIDA, Novartis, Pharmacia & Upjohn, Predix Pharmaceuticals (Epix), and Solvay. TDC is a consultant to the Los Angeles Department of Mental Health on the implementation of early detection and intervention services for youth at risk for psychosis, a consultant to Boehringer Ingelheim Pharmaceuticals, and a co-inventor on a pending patent "Compositions and Methods for Blood Biomarker Analysis for Predicting Psychosis Risk in Persons with Attenuated Psychosis Risk Syndrome." JHK is the editor of *Biological Psychiatry*. JHK consults for AbbVie Inc (formerly Abbott Laboratories), AMGEN, Astellas Pharma Global Development Inc, AstraZeneca Pharmaceuticals, Biomedisyn Corporation, Bristol-Myers Squibb, Eli Lilly and Co, Euthymics Bioscience Inc, and Neurovance Inc, a subsidiary of Euthymic Bioscience, Forum Pharmaceuticals, Janssen Research & Development, Lundbeck Research USA, Novartis Pharma AG, Otsuka America Pharmaceutical Inc, Sage Therapeutics Inc, Sunovion Pharmaceuticals Inc, and Takeda Industries. JHK is on the scientific advisory board for Lohocla Research Corporation, Mnemosyne Pharmaceuticals Inc, Naurex Inc, and Pfizer Pharmaceuticals. JHK holds stock in Biohaven Medical Sciences, and stock options in Mnemosyne Pharmaceuticals Inc. JHK holds three patents/inventions: Seibyl JP, Krystal JH, Charney DS. "Dopamine and noradrenergic reuptake inhibitors in treatment of schizophrenia." US Patent #5,447,948. [September 5, 1995]; Vladimir C, Krystal JH, Sanacora G. "Glutamate Modulating Agents in the Treatment of Mental Disorders" US Patent No. 8,778,979 B2 [July 15, 2014]; Charney D, Krystal JH, Manji H, Matthew S, Zarate C. "Intranasal Administration of Ketamine to Treat Depression" United States Application No. 14/197,767 filed on March 5, 2014; United States application or PCT International application No. 14/306,382 filed on June 17, 2014. PRC was supported by the Connecticut State Department of Mental Health and Addiction Services. PRC was funded by an IMHRO/Janssen Rising Star Translational Research Award and CTSA Grant Number UL1 TR000142 from the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. PRC consults for Otsuka Pharmaceuticals. AMC, ZS, RG, and MKJ declare no competing interests.

Acknowledgments

We thank Gregory McCarthy, Gregory Samanez-Larkin, Jutta Joorman, and Emily Ward for advice and thoughtful comments on the manuscript.

References

1 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:** 870–80.

2 Rost K, Nutting P, Smith JL, Elliott CE, Dickinson M. Managing depression as a chronic disease: a randomised trial of ongoing treatment in primary care. *BMJ* 2002; **325:** 934.

3 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:** 746–58.

4 Cipriani A, Brambilla P, Furukawa T, et al. Fluoxetine versus other types of pharmacotherapy for depression. *Cochrane Database Syst Rev* 2005; **4:** CD004185.

5 Insel TR, Cuthbert BN. Brain disorders? Precisely. *Science* 2015; **348:** 499–500.

6 Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981; **9:** 591–97.

7 Pfeiffer RM, Park Y, Kreimer AR, et al. Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. *PLoS Med 2013;* **10:** e1001492.

8 Wijeysundera DN, Karkouti K, Dupuis JY, et al. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *JAMA* 2007; **297:** 1801–09.

9 Ross PL, Gerigk C, Gonen M, et al. Comparisons of nomograms and urologists' predictions in prostate cancer. *Semin Urol Oncol* 2002; **20:** 82–88.

10 Specht MC, Kattan MW, Gonen M, Fey J, Van Zee KJ. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. *Ann Surg Oncol* 2005; **12:** 654–59.

11 Kattan MW, Yu C, Stephenson AJ, Sartor O, Tombal B. Clinicians versus nomogram: predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology* 2013; **81:** 956–61.

12 Phillips ML, Chase HW, Sheline YI, et al. Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. *Am J Psychiatry* 2015; **172:** 124–38.

13 Uher R, Tansey KE, Rietschel M, et al. Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. *Am J Psychiatry* 2013; **170:** 207–17.

14 Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry* 2013; **74:** 7–14.

15 Odeh MS, Zeiss RA, Huss MT. Cues they use: clinicians' endorsement of risk cues in predictions of dangerousness. *Behav Sci Law* 2006; **24:** 147–56.

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

17 Uher R, Perlis RH, Henigsberg N, et al. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychol Med* 2012; **42:** 967–80.

18 Kuhn M, Johnson K. Applied predictive modeling. London: Springer, 2013.

19 Hastie T, Tibshirani R, J Friedman J. The elements of statistical learning. *Elements* 2009; **1:** 337–87.

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:** 449–59.

21 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:** 689–701.

22 Zimmerman M, Mattia JI. A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. *Arch Gen Psychiatry* 2001; **58:** 787–94.

23 Draper NR, Smith H, Pownell E. Applied regression analysis, vol 3. New York: Wiley, 1996.

24 Berk RA. Regression analysis: a constructive critique. London: Sage, 2004.

25 Kessler RC, Warner CH, Ivany C, et al. Predicting suicides after psychiatric hospitalization in US army soldiers: the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry* 2014; **72:** 49–57.

26 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Statist Soc B 2005;* **67:** 301–20.

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

28 Friedman JH. Stochastic gradient boosting. 1999. https://statweb.stanford.edu/~jhf/ftp/stobst.pdf (accessed Jan 4, 2016).

29 Friedman JH. Recent advances in predictive (machine) learning. *J Classif* 2006; **23:** 175–97.

30 de Ligt J, Willemsen JH, van Bon BWM, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012; **367:** 1921–29.

31 Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002; **38:** 367–78.

32 Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2006; **28:** 1–26.

33 Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods* 2015; **250:** 126–36.

34 Kumbhani DJ, Wells BJ, Lincoff LM, et al. Predictive models for short- and long-term adverse outcomes following discharge in a contemporary population with acute coronary syndromes. *Am J Cardiovasc Dis* 2013; **3:** 39–52.

35 Tiong HJ, Goldfarb DA, Kattan MW, et al. Nomograms for predicting graft function and survival in living donor kidney transplantation based on the UNOS registry. *J Urol* 2009; **181:** 1248–55.

36 Nam RK, Toi A, Klotz LH, et al. Assessing individual risk for prostate cancer. *J Clin Oncol* 2007; **25:** 3582–88.

37 Brennan MF, Kattan MW, Klimstra D, Conlon K. Prognostic nomogram for patients undergoing resection for adenocarcinoma of the pancreas. *Ann Surg* 2004; **240:** 293–98.

38 Leuchter AF, Cook IA, Marangell LB, et al. Comparative effectiveness of biomarkers and clinical indicators for predicting outcomes of SSRI treatment in major depressive disorder: results of the BRITE-MD study. *Psychiatry Res* 2009; **169:** 124–31.

39 Leuchter AF, Cook IA, Hamilton SP, et al. Biomarkers to predict antidepressant response. *Curr Psychiatry Rep* 2010; **12:** 553–62.

40 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.

41 Chekroud AM, Krystal JH. Personalised pharmacotherapy: an interim solution for antidepressant treatment? *BMJ* 2015; **350:** h2502.

42 Paulus MP. Pragmatism instead of mechanism a call for impactful biological psychiatry. *JAMA Psychiatry* 2015; **72:** 631–32.

43 Cuijpers P, Cristea IA. What if a placebo effect explained all the activity of depression treatments? *World Psychiatry* 2015; **14:** 310–11.

44 DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, Lorenzo-Luaces L. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One* 2014; **9:** 1–8.