Received: 3 December 2020 | Revised: 28 January 2021 | Accepted: 16 February 2021

DOI: 10.1111/cts.13035

ARTICLE

Application of machine learning to predict reduction in total PANSS score and enrich enrollment in schizophrenia clinical trials

Jagdeep T. Podichetty[1] | Rebecca M. Silvola[1] | Violeta Rodriguez-Romero[1] | Richard F. Bergstrom[1] | Majid Vakilynejad[2] | Robert R. Bies[1,3,4] | Robert E. Stratford Jr.[1]

[1]Division of Clinical Pharmacology, Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA

[2]Takeda Pharmaceuticals U.S.A., Inc, Cambridge, Massachusetts, USA

[3]Department of Pharmaceutical Sciences, University at Buffalo, State University of New York, Buffalo, New York, USA

[4]Institute for Computational Data Science, University at Buffalo, State University of New York at Buffalo, Buffalo, New York, USA

Correspondence
Robert E. Stratford Jr., Research II, 950 W. Walnut Street, Indianapolis, IN 46202, USA.
Email: robstrat@iu.edu

Funding information
Research reported in this publication was supported by the National Institutes of Health (NIH)/NIGMS T32GM842528 (RMS) in addition to funding obtained from Takeda Pharmaceutical International Co.

Abstract

Clinical trial efficiency, defined as facilitating patient enrollment, and reducing the time to reach safety and efficacy decision points, is a critical driving factor for making improvements in therapeutic development. The present work evaluated a machine learning (ML) approach to improve phase II or proof-of-concept trials designed to address unmet medical needs in treating schizophrenia. Diagnostic data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) trial were used to develop a binary classification ML model predicting individual patient response as either "improvement," defined as greater than 20% reduction in total Positive and Negative Syndrome Scale (PANSS) score, or "no improvement," defined as an inadequate treatment response (<20% reduction in total PANSS). A random forest algorithm performed best relative to other tree-based approaches in model ability to classify patients after 6 months of treatment. Although model ability to identify true positives, a measure of model sensitivity, was poor (<0.2), its specificity, true negative rate, was high (0.948). A second model, adapted from the first, was subsequently applied as a proof-of-concept for the ML approach to supplement trial enrollment by identifying patients not expected to improve based on their baseline diagnostic scores. In three virtual trials applying this screening approach, the percentage of patients predicted to improve ranged from 46% to 48%, consistently approximately double the CATIE response rate of 22%. These results show the promising application of ML to improve clinical trial efficiency and, as such, ML models merit further consideration and development.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

In silico approaches for a priori prediction of patient therapeutic response in a clinical trial could improve clinical trial efficiency by reducing patient enrollment requirements and potentially decision time.

Robert R. Bies and Robert E. Stratford Jr. are co-senior authors.

WHAT QUESTION DID THIS STUDY ADDRESS?

Application of a machine learning (ML) approach was investigated to determine if previously collected, patient-specific data could be used to predict and categorize individual patient treatment response during a clinical trial assessing treatment efficacy in schizophrenia.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

Based on three virtual trials, model application resulted in "improvement" predictions ranging from 46% to 48% compared to actual improvement of 22% in the CATIE trial.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

ML shows promise as a useful tool to supplement patient enrollment to thereby improve clinical trial efficiency.

INTRODUCTION

Schizophrenia affects ~ 1% of the global population.[1] Typical onset during late adolescence to early adulthood, symptom severity, and disease progression combine to result in potential high personal, family, and societal impact.[2] Schizophrenia presents as a cluster of symptoms in affected individuals, including positive symptoms, negative symptoms, and cognitive impairment. Examples of positive symptoms include delusions and hallucinations; negative symptoms encompass decreased affect, emotional withdrawal, and inability to experience pleasure; cognitive impairment reflects deficits in attention, memory, reasoning, and processing speed. The introduction of antipsychotic medications in the 1950s represented a therapeutic breakthrough for treatment of positive symptoms. Further therapeutic improvements, such as reduced incidence of tardive dyskinesia and other adverse side effects, were achieved in the late 1980s with the second-generation antipsychotics, commonly referred to as atypicals. Unfortunately, expansion of the antipsychotic formulary has been slow despite significant research efforts to improve understanding of this disease and identify new treatment modalities. There remains a large unmet medical need for the discovery and development of new therapeutics that retain the safety and efficacy achieved by currently available drugs, but specifically target the negative symptoms and cognitive deficits associated with schizophrenia.[3]

Against this backdrop of lacking novel antipsychotic medication discovery, the costs of new drug development overall have increased dramatically. A report by the Tufts Center for the Study of Drug Development estimated the cost of bringing a new drug to market was $2.6 billion in 2013, a 145% increase from 2003.[4] The large attrition of drug candidates during clinical trials is a major contributor to the growing expense for the development of central nervous system (CNS) targeted drugs; success rates are poor and fall below 10%.[5] Thus, there is great need for more accurate and efficient processes that facilitate rapid clinical testing of potential CNS drugs.

An innovative example that addresses this need is the use of brain magnetic resonance imaging (MRI) to identify anatomic and connectivity abnormalities in patients with schizophrenia[6–8] as a prognostic tool regarding the clinical course of schizophrenia.[9] Furthermore, coupling MRI technology with machine learning (ML) analysis has shown promise to detect CNS diseases even at their earliest manifestation.[10–15] With respect to drug treatment, two longitudinal studies paired functional MRI with ML to predict treatment response to atypical antipsychotics based on connectivity changes in striatal[16] and cortical regions.[17] These studies evaluated the application of ML and MRI as a diagnostic, prognostic, and/or treatment response biomarker to screen patients for enrollment in clinical trials studying treatment efficacy in schizophrenia. Potential downsides of combining MRI with ML include the expense and inconvenience of repeated MRI scans, the limited availability of MRI equipment, and skilled practitioners of these imaging technologies.

To bypass the above concerns, an alternative to using MRI with ML as a treatment response biomarker is the development of ML models that predict or differentiate patient-specific treatment responses based solely on clinical assessment(s). Two of the most commonly used clinical assessment instruments for schizophrenia are the Brief Psychiatric Rating Scale (BPRS)[18,19] and the Positive and Negative Syndrome Scale (PANSS),[20] both of which are used extensively to assess disease severity and antipsychotic treatment efficacy. Modeling of the data from these scales has proven to be valuable, such as the research reported by Krekels et al., in which PANSS scores were modeled over time to successfully differentiate paliperidone versus placebo responses.[21] Congruent with this approach, we hypothesized that an ML model could be developed and used to identify patients who are more likely to experience an efficacious response to antipsychotic therapy using clinical diagnostic

data. To develop this ML model, the CATIE schizophrenia trial[22] was chosen for its well-structured data and detailed assessment of clinical and functional measures of disease severity and patient response.[23] Briefly, CATIE compared the efficacy and safety of atypical antipsychotic drugs: olanzapine, quetiapine, risperidone, clozapine, ziprasidone, to the typical antipsychotics perphenazine and fluphenazine decanoate. Among many clinical studies, CATIE is one of the most comprehensive and data-rich independent trials to examine existing therapies for schizophrenia.[24]

METHODS

Data collection and preparation

De-identified individual patient level clinical data from the CATIE trial were requested and obtained from the trial sponsor, the National Institute of Mental Health (NIMH). In addition, data from two other trials evaluating safety and efficacy of approved antipsychotics: A Comparison of Long-acting Injectable Medications for Schizophrenia (ACLAIMS)[25] and Preventing Relapse in Schizophrenia: Oral Antipsychotics Compared to Injectables: Evaluating Efficacy (PROACTIVE)[26] were obtained. For all three databases, the Indiana University Institutional Review Board reviewed and provided an exemption to support this research project.

The CATIE trial, conducted from October 2001 to December 2004, established regularly scheduled patient assessments, and collected data for up to 18 months. Targeted study population enrollment was 1600 patients with collection of up to 500 attributes for each individual patient. Eligibility requirements included participants aged between 18 and 65 years and a previous diagnosis of schizophrenia.[24] Exclusion criteria included other cognitive disorders, such as schizoaffective disorder, mental retardation, pervasive developmental disorder, delirium, dementia, amnesia; a history of serious adverse reaction to the proposed medication; a history of only one schizophrenic episode; a history of treatment resistance; and women currently pregnant or breast-feeding. Of note, patients with tardive dyskinesia were eligible to enroll but were restricted from assignment to perphenazine. At baseline, the incidence of comorbidities within the study population was as follows: 11% diabetes, 14% hyperlipidemia, and 20% hypertension.[24] We did not evaluate potential impact of these comorbidities on treatment response due to absence of comorbidity data collected over the course of the trial. Additionally, whereas we acknowledge potential pharmacokinetic interactions between the antipsychotic agents of study and any permitted concomitant medications, analysis for such interactions was not included in the model's development due to dataset limitations. Finally, an implicit assumption was that drug-specific therapeutic steady-state pharmacokinetics and stable disease applied for the duration of on-treatment assessments. For example, factors influencing target site drug concentration, such as protein binding, liver, and/or renal function, were assumed stable.

The CATIE dataset, like most clinical trial data, came from the NIMH as a set of individual files containing different patient attributes. The longitudinal data files required careful examination for measurement dates and/or missing values, as patient attributes were often measured at inconsistent time intervals or frequencies. A 2-week measurement window was used to approximate time of measurement and the recorded "visit day" was converted from days to months and then rounded to the closest whole number. Data for months 2, 4, and 5 were not included, as observations were too infrequent to contribute to the ML model identification. All patient attribute data were combined based on patient ID, with each row representing one patient and columns representing responses over time for each diagnostic instrument.

The following paragraph describes the process used to create the curated dataset. Of the 1894 subjects screened in the CATIE trial (actual enrollment was greater than the target enrollment of 1600), 434 subjects were excluded by CATIE trial authors due to concerns about the integrity of the data.[24] Of the 1460 subjects remaining, 658 did not have recorded outcomes at 6 months and were excluded. Finally, with interest in improvement at 6 months in mind and 802 subjects remaining, 163 subjects who had already experienced clinical improvement at 3 months were excluded. This was done to avoid bias in model training as data from these 163 subjects, baseline and improved PANSS scores at 3 months, would have biased the model toward predicting improvement for similar baseline and 3 months PANSS scores without considering other features.

Subsequent to the initial work described in the preceding paragraph, the CATIE dataset was organized and formatted to support ML model development by using the R package, dplyr.[27] Assessments from the following diagnostic evaluations were included in the curated dataset: PANSS, Clinical Global Impression of Severity (CGI), quality of life, structured clinical interview of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, neurocognitive battery, vital signs, and Calgary Depression Scale for Schizophrenia (CDSS). The various items and/or subscales within each of these instruments were considered. When more than 70% of the values for any given item were missing, that item was excluded. Imputations were not conducted for any given item. Data were categorized at baseline, 1 month, and 3 months of treatment for each patient. Based on the work of others that trajectories in PANSS scores over time correlate with relapse,[28] slopes of PANSS subscales (general, positive symptoms, and negative symptoms) over time were calculated per patient and assessed for ability to improve ML model performance.

A binary classification model ML approach was used such that a positive response to therapy, "improvement," was defined as 20% or more reduction in total PANSS score between baseline and 6 months; whereas "no improvement" was defined as an inadequate treatment response (<20% reduction in total PANSS). The PANSS rating scale was chosen as the objective response variable for its established use as a standard assessment of clinical efficacy in the treatment of schizophrenia.[17,29,30] When using percent reduction in the PANSS score to identity treatment response, standard cutoffs typically range from 20% to 50%.[31] We chose a cutoff of 20% to develop a model able to detect modest, albeit clinically justifiable, improvement.[32]

Full ML model development

The present work aimed to develop a binary classification ML model trained to predict treatment response (specifically, a 20% reduction in total PANSS score) at 6 months based upon patient responses at baseline, 1 month, and 3 months, assuming that antipsychotic exposure was consistently within therapeutic range. Although the CATIE trial collected data up to 18 months, we found that data beyond 6 months was too sparse to inform model training. Several classification algorithms were tested and compared. These were random forest (RF), logistic regression, naïve Bayes, and support vector machine. The RF algorithm, a tree-based classifier consisting of multiple individual decision trees, each generated by randomly subsampling the training data,[33] was ultimately chosen, as it had the best overall performance with respect to receiver operating characteristic (ROC) curve, true positive rate (TPR), true negative rate (TNR), and correct classification rate (CCR), the latter three as defined below in Equations 1–3, respectively.

Figure 1 depicts the workflow developed to build the binary, RF ML classification model (full ML model) to categorize each patient's response as either "improvement" (20% or more reduction in total PANSS score between baseline and 6 months) or "no improvement" using the curated dataset (639 patients). Detailed description of the workflow process is found in Supplementary Methods.

Several classification algorithms were then tested and compared. These included RF, logistic regression, naïve Bayes, and support vector machine. An ensemble modeling approach was adopted using the RF algorithm, wherein a number of different decision trees were used to make predictions.[34] Model training was performed using 5, 7, and 10-fold cross-validation to avoid overfitting.[35,36] Internal model validation, step 5, was conducted using the testing dataset ($n = 192$) created in step 2.

Following model development, the full ML model's 3-month response predictions using the testing dataset were compared to CATIE outcomes. Model accuracy was assessed using the ROC curve. Accuracy is defined as area under the ROC curve (ROC AUC), which relates TPR (model sensitivity, plotted on the y-axis), and TNR (model selectivity, or sometimes referred to as specificity, plotted on the *x*-axis). The equations used to calculate TPR, TNR, and CCR are shown below.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{1}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{2}$$

$$\mathrm{CCR} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP})} \tag{3}$$

Application of ML approach for clinical trial enrichment

To evaluate if the full ML model developed could be used to identify patients most likely to exhibit clinical improvement with antipsychotic therapy, a second ML model was developed. The concept was to determine if an ML model provided with baseline data alone could support recruitment and provide feedback relatively quickly (3 months instead of 6 months) to identify patients most likely to experience an "improvement" in response to the antipsychotics of study. Thus, this so-called patient screening ML model differed from the full ML model in that only baseline data (no data at 1 and 3 months) were used to predict outcome at 3 months. The same workflow illustrated in Figure 1 was used to develop this screening model. However, patient numbers used to develop the screening model ($n = 1009$ before splitting into training and testing datasets) were larger than the first model ($n = 639$) because fewer patients were providing data as study duration reached 6 months. The screening model trade-off, in using only baseline data, meant the dataset, although larger in individual patient numbers, did not contain multiple measurements collected over time. Figure S1 compares patient numbers used for the two models.

To assess performance of the patient screening ML model, the workflow shown in Figure 2 was used to conduct a series of virtual clinical trials consisting of preselected actual patients from the CATIE trial. Specifically, patients predicted not to improve at 3 months were excluded from enrollment. Three such virtual clinical trials were conducted, all of 3 months' duration and consisting of 50 patients each. The percentage of patients predicted to improve at month 3 from each virtual trial was then compared with actual CATIE outcomes at 3 months.

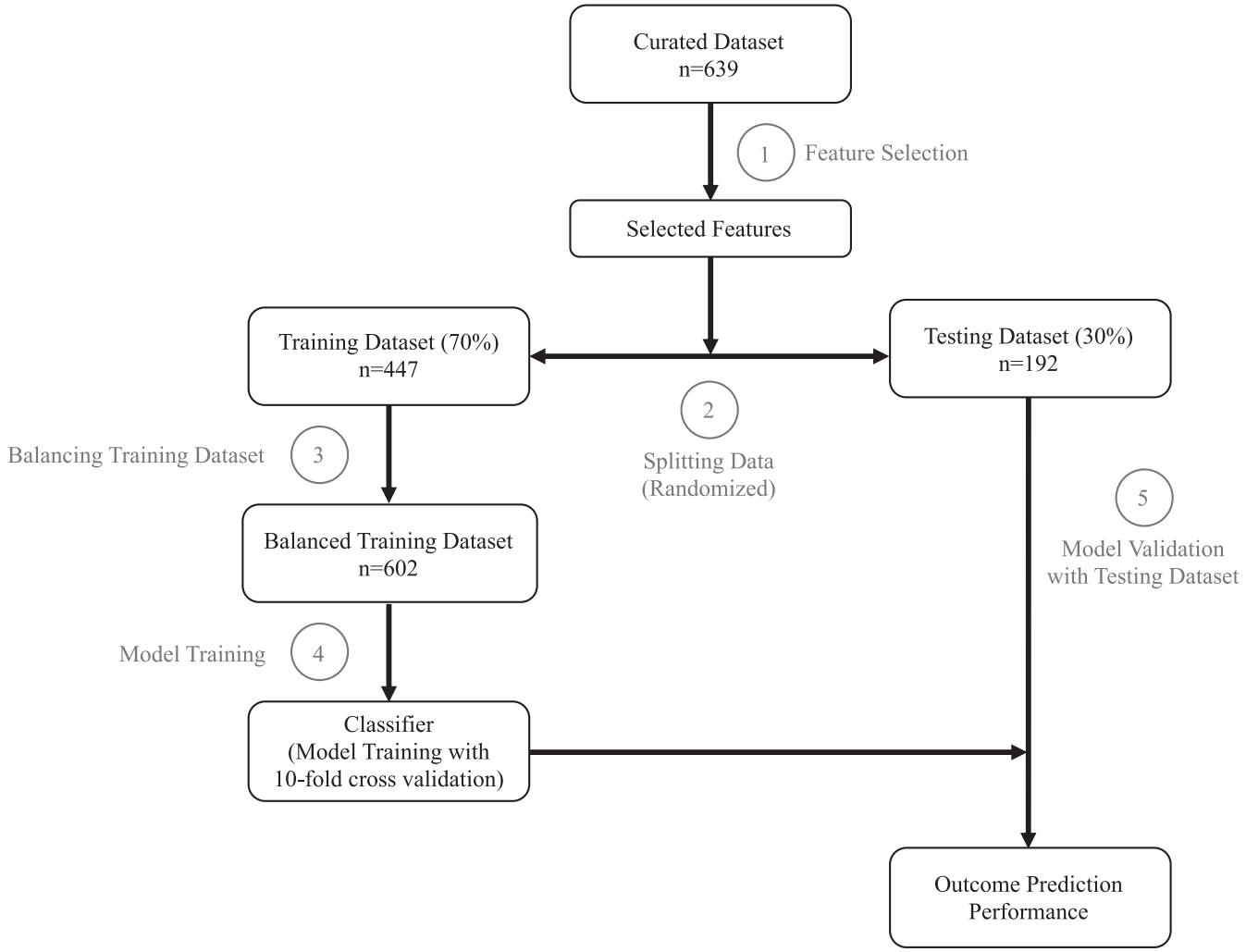

FIGURE 1 Five-step workflow for full machine learning (ML) model analysis to develop a binary classification model to predict treatment response outcome based on clinical measurements taken at baseline, 1 month, and 3 months

RESULTS

Data curation for ML analysis

The final dataset used for the full ML model developed according to the scheme outlined in Figure 1 consisted of 639 patients and 397 features total. Baseline patient demographic and clinical characteristics are summarized in Table S1. Figures S2–S6 illustrate distributions of baseline measurements for several of the diagnostic evaluation instruments applied in CATIE. Patient classification as "improvement" versus "no improvement" came from data at 6 months. Of all 397 features, a subset of 123 features (Table S2) was selected using Pearson's correlation to provide a final patient/feature ratio of five. Following feature selection, the dataset contained more patients in the "no improvement" category (Figure 3a). This imbalance was preserved after the random split into separate training and testing datasets (Figure 3b,c). After balancing the training set, similar proportions of patients were obtained in both categories for proper model training (Figure 3d).

Model performance

Among the several algorithms evaluated, the RF algorithm had the best overall performance. For this algorithm, model performance in the test stage yielded an accuracy of 0.7 (ROC AUC), which was considered reasonable. In contrast, accuracy for the other algorithms ranged from 0.58 to 0.65. Table 1 summarizes the RF performance for the full ML model training and testing stages with respect to ROC, TPR, TNR, and CCR. Poor performance with respect to TPR was likely due to the low number of patients in the "improvement" category within the test stage dataset. However, the high TNR and CCR values in the test stage demonstrate model ability to identify patients with "no improvement."

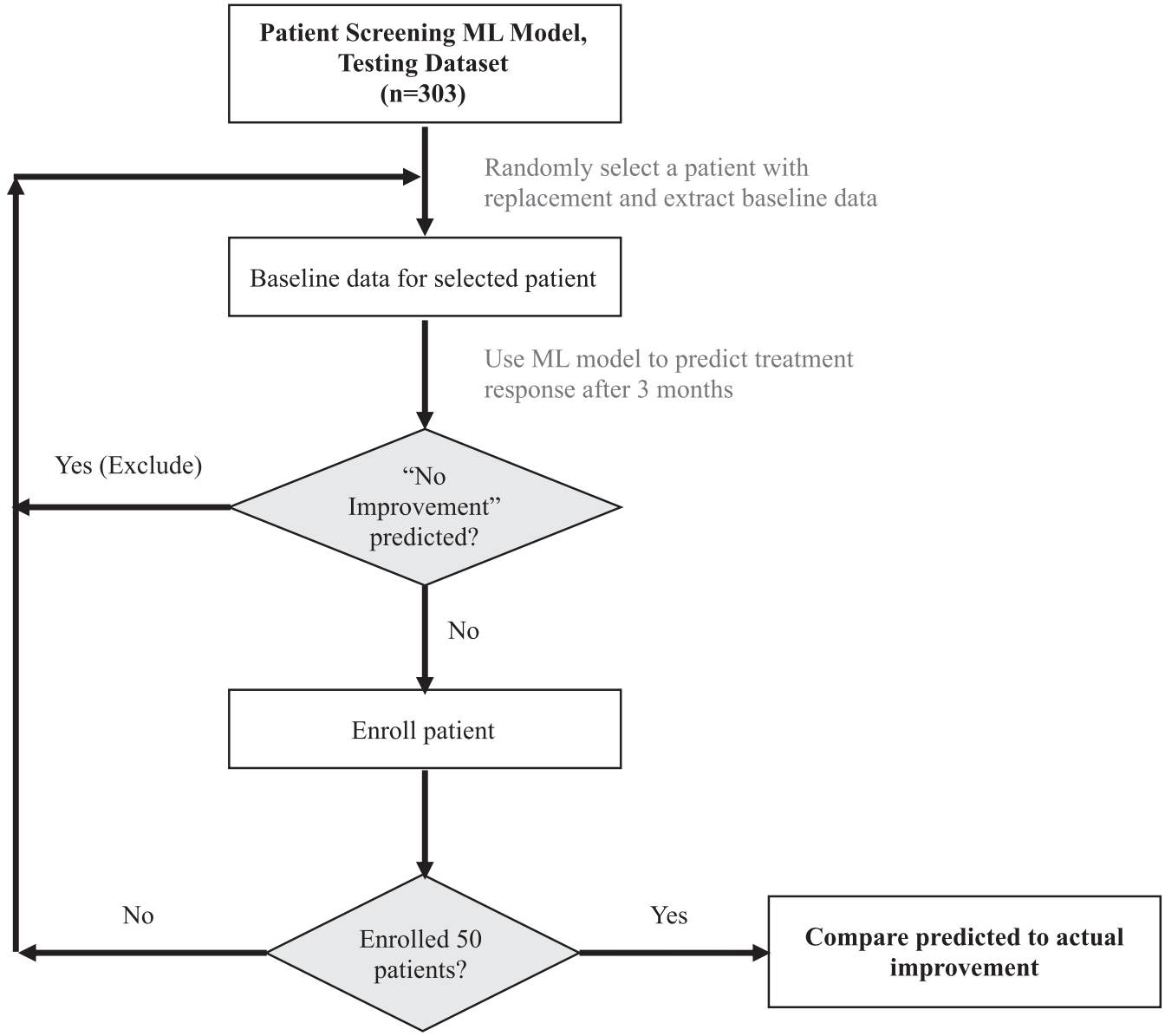

FIGURE 2 Schematic of workflow to assess patient screening machine learning (ML) model ability to enhance clinical trial population enrollment for the virtual trials

As indicated in Figure 4, 8 of the 10 patient attributes with the highest contribution to predict patient response came from PANSS subscales measured at baseline.

Clinical trial enrichment

In contrast to poorly predicting patients in the "improvement" category, the patient screening ML model performed reasonably well at predicting "no improvement" at 3 months (Table 2). Application of this model to the flow scheme specified in Figure 2 to enrich patient selection by excluding patients predicted not to improve led to consistent predictions of more than 45% of patients improving with therapy after 3 months (Table 3). These results compare to 22% of patients that actually experienced improvement in the CATIE trial. Increased ability to identify patients with a greater propensity to be in the "improvement" category was attributed to high model specificity (TNR).

DISCUSSION

Data curation for ML analysis

Selection of data for which an ML model will be developed is a critical component of success. In addition to the

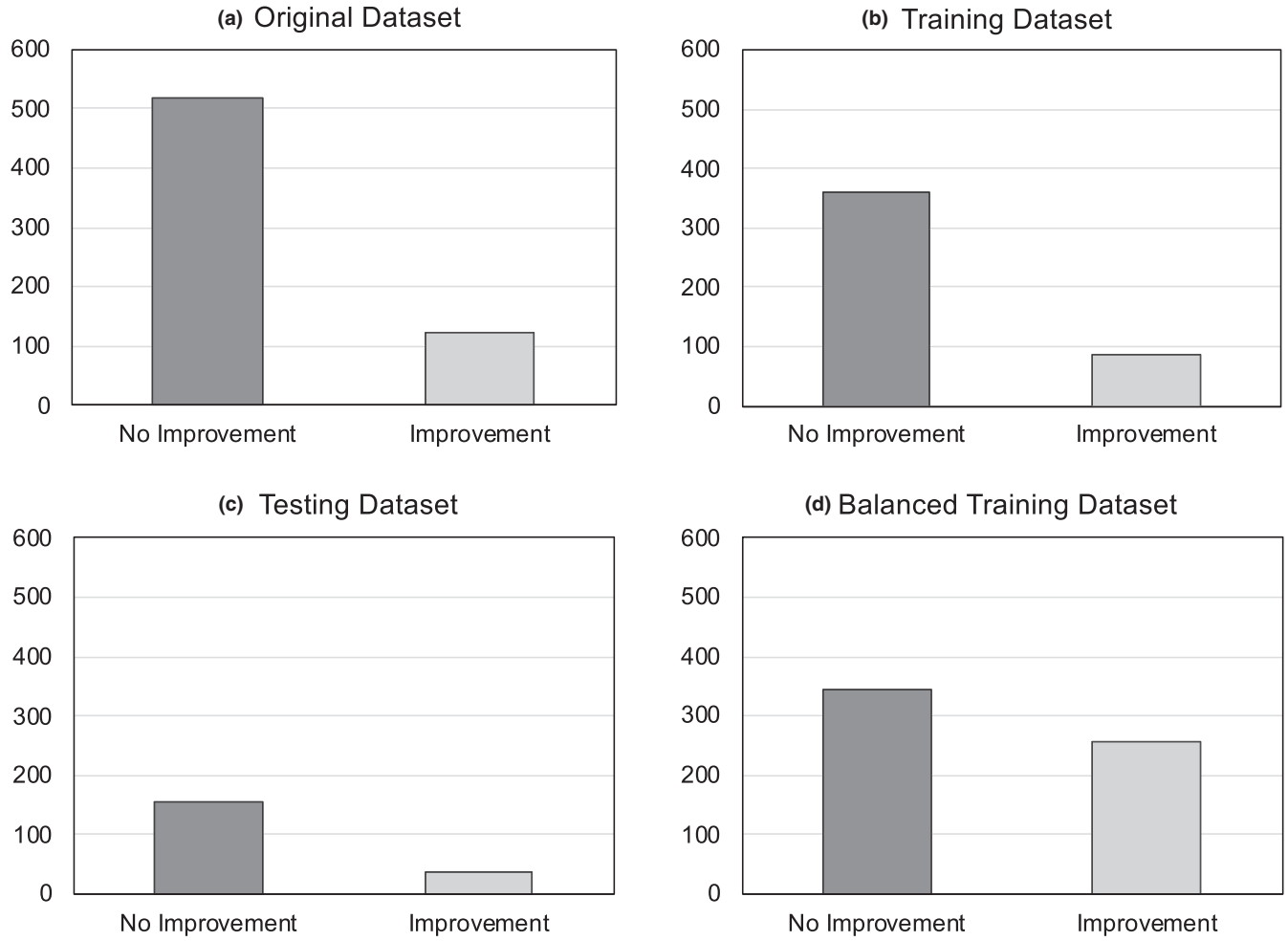

FIGURE 3 Patient distribution between "no improvement" and "improvement" categories of the (a) original curated dataset, $n = 639$; (b) training dataset, $n = 447$; (c) testing dataset, $n = 192$; and (d) balanced training dataset, $n = 602$

TABLE 1 Performance of the full ML model during training and testing stages

RF	ROC	TPR (sensitivity)	TNR (specificity)	CCR	TP	FP	TN	FN
Training	0.956	0.740	0.991	0.884	191	3	341	67
Testing	0.700	0.194	0.936	0.800	7	10	146	29

Abbreviations: CCR, correct classification rate; FN, false negative; FP, false positive; ML, machine learning; RF, random forest; ROC, receiver operator characteristic; TN, true negative; TNR, true negative rate; TP, true positive; TPR, true positive rate.

CATIE Schizophrenia trial, there are at least two other publicly available datasets from phase IV clinical trials designed to prospectively evaluate safety and efficacy of antipsychotics used for the treatment of schizophrenia: ACLAIMS[25] and PROACTIVE.[26] Attempts to develop an ML model with data from the ACLAIMS trial fell below suitable expectations. Specifically, the dataset yielded ROC results less than 0.6 in both training and testing stages of ML model development. This poor performance was attributed to the smaller trial population ($n = 311$) and the collection of patient features, which, collectively, were not informative to ML model development. Interestingly, none of the collected features were consistent with the top 10 features identified from the CATIE trial (Figure 4). Due to comparably small study size ($n = 357$), no attempt was made to develop an ML model from the PROACTIVE trial data. The CATIE clinical trial dataset included rich feature data from multiple cognitive and behavioral assessment instruments and a large number of subjects, making it well-suited to support ML model development. On the

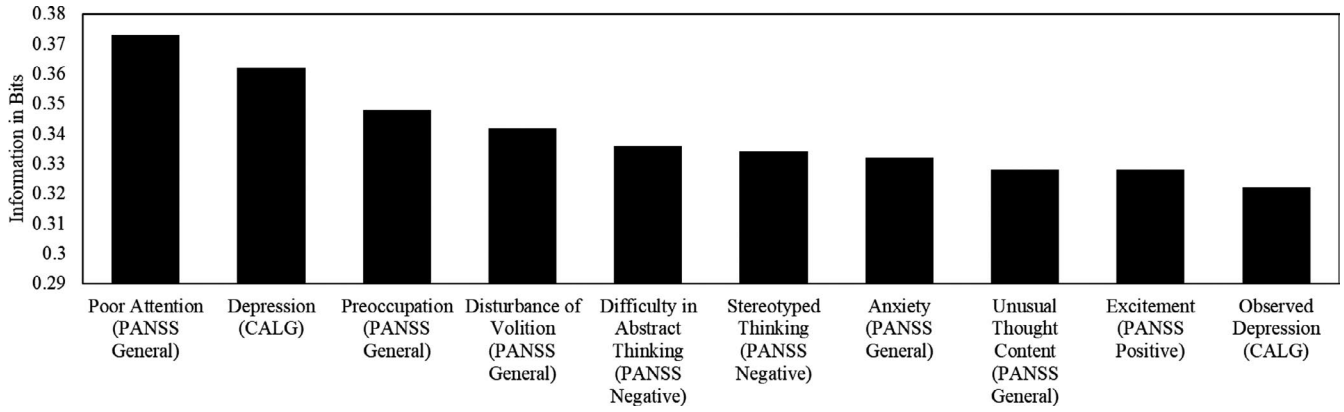

FIGURE 4 Top 10 patient attributes most predictive of treatment response outcome. CALG, Calgary; PANSS, Positive and Negative Syndrome Scale

TABLE 2 Performance of patient screening ML model during training and testing stages using only baseline data

RF	ROC	TPR (sensitivity)	TNR (specificity)	CCR	TP	FP	TN	FN
Training	0.956	0.714	0.985	0.869	332	9	611	133
Testing	0.653	0.167	0.948	0.762	12	12	219	60

Abbreviations: CCR, correct classification rate; FN, false negative; FP, false positive; ML, machine learning; RF, random forest; ROC, receiver operator characteristic; TN, true negative; TNR, true negative rate; TP, true positive; TPR, true positive rate.

TABLE 3 Patient screening ML model performance during the virtual clinical trials

Virtual clinical trials	Number of patients enrolled	Number of patients predicted to improve at 3 months	Percentage of patients predicted to improve at 3 months
1	50	24	48
2	50	23	46
3	50	24	48

Abbreviation: ML, machine learning.

other hand, although the CATIE trial collected data up to 18 months, we found that data beyond 6 months was too sparse to inform model training. This study demonstrates the potential of ML techniques to both query the study data for predictive relationships in novel and unstructured ways while providing a potential pathway to enrich future studies by deploying strategies that can enrich the patient population. Yet, more work remains to be done in the application of these modeling techniques. Although ML models may prove to be applicable at any stage of drug development, our present interest was to evaluate them in the early stages, where proof-of-concept is the goal, rather than in definitive or confirmatory phase III studies.

Universally, in drug development, there is a clear advantage to leveraging opportunities in trial designs that permit a shorter trial duration.[37] Studies that require longitudinal patient-level data of 6 months' duration or more are too long to facilitate the general framework of a quick win/quick kill proof-of-concept strategy.[38,39] However, diseases like schizophrenia often require studies of a longer duration to achieve meaningful changes in markers of efficacy and safety. To address these challenges, the probability of successfully predicting improvement at 18 months, using data from the first 6 months, was evaluated. To our disappointment, this approach did not meet performance requirements (ROC AUC was $\leq$0.7), which was attributed to an insufficient number of "improvement" cases at 18 months to support model training. Ideally, baseline information alone would be sufficient to predict improvement at 3 or 6 months. Addition of 1-month and 3-month data to baseline scores was necessary to properly train the full ML model to achieve ROC AUC greater than or equal to 0.7 when predicting treatment response at 6 months. Addition of slopes describing change in each of the PANSS subscales (general, positive, and negative) from

baseline to 1 month and baseline to 3 months also improved accuracy of the full ML model when predicting outcome at 6 months.

ML model development and performance

As no single algorithm fits all ML models, several were evaluated, including logistic regression, naïve Bayes, and support vector machine. The RF algorithm had the best overall performance according to the criteria specified in Methods and shown in Table 1. A tree-based ML classification algorithm, such as RF, offers several advantages over other parametric models, such as logistic regression.[40] First, decision trees do not require a specific relationship between covariates and outcome. Second, these algorithms require less data preparation, as features do not require scaling or centering prior to model training. Tree-based ML classification also requires fewer assumptions as it has no functional form. On the other hand, the approach can be susceptible to overfitting and may require regularization.[41–44] Regularization requires adding constraints to simplify the model and curb risk of overfitting. Outcome prediction from RF is based on the most common outcomes from individual decision trees.

Model accuracy was assessed using ROC AUC. As binary classification models frequently perform better at classifying one category, it was no surprise when the present model performed significantly better at predicting the "no improvement" category. Although there are several reasons why this would occur, the most obvious is the reduced number patients who met the "improvement" definition. In other words, there was limited availability of outcomes from the "improvement" group. Two additional possibilities are high noise and low signal.

To optimize model performance, various split ratios for the training and testing populations, as well as different cross validation folds were explored. Split ratios of 70:30, 75:25, and 80:20 were tested with the 70:30 ratio yielding the highest number of subjects in the "improvement" category. We tested 5, 7, and 10-fold cross-validation folds and found the curated dataset able to support the more rigorous 10-fold cross-validation. All splits and cross-validations were randomized. As described previously, the training dataset was balanced using the SMOTE function in R software to preserve and enhance the number of "improvement" cases and to restrain sampling of "no improvement" cases. Although the RF algorithm showed a TPR greater than 0.7 during the training stage, this measure of sensitivity of model performance was reduced in the testing dataset to 0.194. This was attributed to the limited number of "improvement" cases. Despite low sensitivity, ROC and CCR values were within a desired range and the model performed well regarding its selectivity, that is, in predicting patients who did not experience clinical improvement at 6 months (TNR = 0.936; Table 1). Hence, the ML model demonstrated potential for use as a tool to identify patients for inclusion, which could be used as a technique to enrich enrollment.

Clinical trial enrichment using ML

Patient recruitment can be a challenging limitation to successful trials. Clinical trials in schizophrenia are no exception. In response to these concerns, the high selectivity of the full ML model was leveraged to develop the patient screening ML model that could cull potential subjects during the enrollment phase not likely to experience an efficacious response. This screening model was created by limiting clinical measurements to those collected only at baseline to predict outcome at 3 months, which is reasonably short and therefore attractive from the standpoint of making quick win/quick kill decisions. Based on higher predicted efficacious response rates (>45%) in 3 virtual trials of 50 patients each as compared with actual results (22%), these findings offer encouragement regarding further investigation of this approach. Although the false-positive rate was low in this screening model (1–0.948 from Table 2), it was not 100% selective with respect to false-positive removal. Had this been achieved, higher predicted response rates would likely have been realized. From the standpoint of efficiency, this potential benefit for a shorter treatment phase needs to be weighed against the potential for a longer recruitment phase in order to identify a sufficient number of patients not predicted to be in the nonresponder group. Additionally, incorporation of an ML model to support patient recruitment may have merit as a novel approach to support patient segmentation in clinical trial design, which may enhance signal detection of an effect from an investigated novel therapy.

It is important to emphasize that the present analysis for ML utility for clinical trial enrollment enrichment in schizophrenia trials is a proof-of-concept; further investigation is necessary. A potential limitation of this specific ML model approach is that the model was trained on clinical responses to antipsychotic drugs that are presumed to be effective based upon their pharmacology at the dopamine and serotonin receptors. These are the common mechanisms relevant to the agents tested in CATIE. Accordingly, it is likely that this model would be more relevant for new chemical entities seeking to improve treatment efficacy and/or safety through similar mechanisms (for example, through increased receptor occupancy or potency, improved receptor selectivity, or enhanced pharmacokinetic properties). Therefore, we conclude that the concept of incorporating an ML-based patient screening approach, as developed herein, merits further research. Furthermore, given the encouraging results obtained, the work merits an expansion of ML techniques to predict the

response of existing and future CNS agents based on a variety of mechanisms in other patient populations, such as in major depression, attention-deficit disorder, pain, Alzheimer's disease, and cognitive impairment, to name a few.

A second limitation of this approach is that features identified represent a particular patient population, and that this population is not going to be adequately represented with respect to potential for response in the space where cognitive issues and negative symptoms, which remain a significant unmet need in medical treatment of schizophrenia, are the primary focus. Finally, an ML approach provides the opportunity to consider a much higher dimensionality of potential input variables to predict outcome compared to a linear or nonlinear mixed effects model. Despite this apparent advantage, there is significant risk of overfitting with ML based approaches. Furthermore, ML approaches do not consider mechanistic underpinnings of a system; therefore, if the system changes (or the training data contain sufficiently heterogeneous responses), the ML algorithm is less likely to predict an outcome correctly, whereas a mixed effects approach (or quantitative systems pharmacology potentially) could accommodate those changes given there may be more mechanistic information. As additional information becomes available, in particular longitudinal information, differences in predictive ability of the ML versus mixed effect models may narrow.

In summary, an ML classification model was developed to predict patient treatment response to currently utilized antipsychotic medications. Overall model performance was satisfactory aside from low sensitivity (TPR < 0.7). Model specificity, represented by the false negative rate, was over 93%. This outcome was leveraged to develop a patient screening ML model to recommend exclusion from a schizophrenia trial designed to demonstrate efficacy. A proof-of-concept analysis via 3 virtual trials of 50 patients each predicted an average 47% of patients would improve at 12 weeks compared to 22% from CATIE trial results.

ACKNOWLEDGEMENTS

This research utilized data from the CATIE clinical trial[22] obtained from NIMH. The opinions or views of the authors in this manuscript does not reflect opinions and views of NIMH.

CONFLICT OF INTEREST

R.B. receives funding as an expert consultant through Belmore Neidrauer LLP for Janssen Pharmaceuticals. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

J.T.P., R.M.S., V.R.R., R.F.B., R.R.B., and R.E.S. wrote the manuscript. M.V., R.F.B., and R.R.B. designed the research. J.T.P., V.R.R., R.F.B., R.R.B., and R.E.S. performed the research. J.T.P. analyzed the data.

ORCID

Rebecca M. Silvola https://orcid.org/0000-0003-0493-0603
Violeta Rodriguez-Romero https://orcid.org/0000-0002-4918-0984
Robert R. Bies https://orcid.org/0000-0003-3818-2252

REFERENCES

1. Freedman R. Schizophrenia. *N Engl J Med*. 2003;349:1738-1749.
2. Cloutier M, Sanon Aigbogun M, Guerin A, et al. The economic burden of schizophrenia in the United States in 2013. *J Clin Psychiatry*. 2016;77:764-771.
3. Patel KR, Cherian J, Gohil K, Atkinson D. Schizophrenia: overview and treatment options. *P & T*. 2014;39:638-645.
4. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ*. 2016;47:20-33.
5. van den Brink WJ, Hankemeier T, van der Graaf PH, de Lange ECM. Bundling arrows: improving translational CNS drug development by integrated PK/PD-metabolomics. *Expert Opin Drug Discov*. 2018;13:539-550.
6. Shenton ME, Dickey CC, Frumin M, McCarley RW. A review of MRI findings in schizophrenia. *Schizophr Res*. 2001;49:1-52.
7. Calhoun VD, Eichele T, Pearlson G. Functional brain networks in schizophrenia: a review. *Front Hum Neurosci*. 2009;3:17.
8. Karlsgodt KH, Sun D, Cannon TD. Structural and functional brain abnormalities in Schizophrenia. *Curr Dir Psychol Sci*. 2010;19:226-231.
9. Li M, Li X, Das TK, et al. Prognostic utility of multivariate morphometry in schizophrenia. *Front Psychiatry*. 2019;10:245.
10. Rathi Y, Malcolm J, Michailovich O, et al. Biomarkers for identifying first-episode schizophrenia patients using diffusion weighted imaging. *Med Image Comput Comput Assist Interv*. 2010;13:657-665.
11. Borgwardt S, Koutsouleris N, Aston J, et al. Distinguishing prodromal from first-episode psychosis using neuroanatomical single-subject pattern recognition. *Schizophr Bull*. 2013;39:1105-1114.
12. Mourao-Miranda J, Reinders AATA, Rocha-Rego V, et al. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychol Med*. 2012;42:1037-1047.
13. Pettersson-Yeo W, Benetti S, Marquand AF, et al. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol Med*. 2013;43:2547-2562.
14. Sun D, van Erp TGM, Thompson PM, et al. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biol Psychiatry*. 2009;66:1055-1060.
15. Schwarz D, Kasparek T. Brain morphometry of MR images for automated classification of first-episode schizophrenia. *Information Fusion*. 2014;19:97-102.
16. Sarpal DK, Robinson DG, Lencz T, et al. Antipsychotic treatment and functional connectivity of the striatum in first-episode schizophrenia. *JAMA Psychiatry*. 2015;72:5-13.
17. Cao B, Cho RY, Chen D, et al. Treatment response prediction and individualized identification of first-episode drug-naive schizophrenia using brain functional connectivity. *Mol Psychiatry*. 2020;25(4):906-913.

18. Flemenbaum A, Zimmermann RL. Inter- and intra-rater reliability of the Brief Psychiatric Rating Scale. *Psychol Rep*. 1973;32:783-792.
19. Yesavage JA. Inpatient violence and the schizophrenic patient. A study of Brief Psychiatric Rating Scale scores and inpatient behavior. *Acta Psychiatr Scand*. 1983;67:353-357.
20. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987;13:261-276.
21. Krekels E, Novakovic AM, Vermeulen AM, Friberg LE, Karlsson MO. Item response theory to quantify longitudinal placebo and paliperidone effects on PANSS scores in schizophrenia. *CPT Pharmacometrics Syst Pharmacol*. 2017;6:543-551.
22. Health. CATIE- Schizophrenia Trial. 2000. https://ClinicalTrials.gov/show/NCT00014001. Accessed April 12, 2021.
23. Stroup TS, McEvoy JP, Swartz MS, et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull*. 2003;29:15-31.
24. Lieberman JA, Stroup TS, McEvoy JP, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med*. 2005;353:1209-1223.
25. Institute, National Institute of Mental Health, Duke University & University of North Carolina, Chapel Hill. A comparison of long-acting injectable medications for schizophrenia. 2011. HYPERLINK "sps:urlprefix::https" https://ClinicalTrials.gov/show/NCT01136772. Accessed April 12, 2021.
26. Health & National Institute of Mental Health. Preventing relapse in schizophrenia: oral antipsychotics compared to injectables: evaluating efficacy. 2006. https://ClinicalTrials.gov/show/NCT00330863. Accessed April 12, 2021.
27. Select Random Samples in R using Dplyr – (sample_n and sample_frac). 2018. https://www.datasciencemadesimple.com/select-random-samples-r-dplyr-sample_n-sample_frac/. Accessed July 5, 2018.
28. Wang D, Gopal S, Baker S, Narayan VA. Trajectories and changes in individual items of positive and negative syndrome scale among schizophrenia patients prior to impending relapse. *NPJ Schizophrenia*. 2018;4 1:10 https://doi.org/10.1038/s41537-018-0056-6
29. Leucht S, Barabássy Á, Laszlovszky I, et al. Linking PANSS negative symptom scores with the Clinical Global Impressions Scale: understanding negative symptom scores in schizophrenia. *Neuropsychopharmacology*. 2019;44:1589-1596.
30. Santor DA, Ascher-Svanum H, Lindenmayer JP, Obenchain RL. Item response analysis of the Positive and Negative Syndrome Scale. *BMC Psychiatry*. 2007;7:66.
31. Leucht S. Measurements of response, remission, and recovery in schizophrenia and examples for their clinical application. *J Clin Psychiatry*. 2014;75(Suppl 1):8-14.
32. Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel RR. What does the PANSS mean? *Schizophr Res*. 2005;79:231-238.
33. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. *IEEE Trans Pattern Anal Mach Intell*. 2007;29:173-180.
34. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
35. Ghojogh B, Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv: 1905.12787. https://arxiv.org/abs/1905.12787. Accessed April 12, 2021.
36. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. New York, NY: Elsevier Inc.; 2011.
37. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med*. 2018;16:29.
38. Bonabeau E, Bodick N, Armstrong RW. A more rational approach to new-product development. *Harv Bus Rev*. 2008;86(3):96-102.
39. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov Today*. 2012;17:1088-1102.
40. Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA. Machine learning: an indispensable tool in bioinformatics. *Methods Mol Biol*. 2010;593:25-48.
41. Kukacka J, Golkov V, Cremers D. Regularization for deep learning: a taxonomy. 2017. https://arxiv.org/pdf/1710.10686v1.pdf. Accessed April 12, 2021.
42. Lemberger PO. On generalization and regularization in deep learning: an introduction for data scientists. arXiv 2017. https://export.arxiv.org/pdf/1704.01312. Accessed April 12, 2021
43. Nusrat I, Jang S-B. A comparison of regularization techniques in deep neural networks. *Symmetry*. 2018;10:648.
44. Raff E, Sylvester J, Mills S. Fair forests: regularized tree induction to minimize model bias. arXiv.org 2017. https://arxiv.org/abs/1712.08197v1. Accessed April 12, 2021

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Podichetty JT, Silvola RM, Rodriguez-Romero V, et al. Application of machine learning to predict reduction in total PANSS score and enrich enrollment in schizophrenia clinical trials. *Clin Transl Sci*. 2021;14:1864–1874. https://doi.org/10.1111/cts.13035