Published in final edited form as:
J Child Psychol Psychiatry. 2022 November ; 63(11): 1347–1358. doi:10.1111/jcpp.13580.

Evidence for machine learning guided early prediction of acute outcomes in the treatment of depressed children and adolescents with antidepressants

Arjun P. Athreya[1], **Jennifer L. Vande Voort**[2], **Julia Shekunov**[2], **Sandra J. Rackley**[2], **Jarrod M. Leffler**[2], **Alastair J. McKean**[2], **Magdalena Romanowicz**[2], **Betsy D. Kennard**[3], **Graham J. Emslie**[3,4], **Taryn Mayes**[3], **Madhukar Trivedi**[3], **Liewei Wang**[1], **Richard M. Weinshilboum**[1], **William V. Bobo**[5], **Paul E. Croarkin**[2]

[1]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN, USA;

[2]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA;

[3]Peter O'Donnell Jr. Brain Institute and the Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, USA;

[4]Children's Health, Children's Medical Center, Dallas, TX, USA;

[5]Department of Psychiatry and Psychology, Mayo Clinic, Jacksonville, FL, USA

Abstract

Background: The treatment of depression in children and adolescents is a substantial public health challenge. This study examined artificial intelligence tools for the prediction of early outcomes in depressed children and adolescents treated with fluoxetine, duloxetine, or placebo.

Methods: The study samples included training datasets ($N = 271$) from patients with major depressive disorder (MDD) treated with fluoxetine and testing datasets from patients with MDD treated with duloxetine ($N = 255$) or placebo ($N = 265$). Treatment trajectories were generated using probabilistic graphical models (PGMs). Unsupervised machine learning identified specific depressive symptom profiles and related thresholds of improvement during acute treatment.

Results: Variation in six depressive symptoms (difficulty having fun, social withdrawal, excessive fatigue, irritability, low self-esteem, and depressed feelings) assessed with the Children's Depression Rating Scale-Revised at 4–6 weeks predicted treatment outcomes with fluoxetine at 10–12 weeks with an average accuracy of 73% in the training dataset. The same six symptoms predicted 10–12 week outcomes at 4–6 weeks in (a) duloxetine testing datasets with an average accuracy of 76% and (b) placebo-treated patients with accuracies of 67%. In

Correspondence: Paul E. Croarkin, Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN 55905, USA; croarkin.paul@mayo.edu.

Supporting information
Additional supporting information may be found online in the Supporting Information section at the end of the article:

Conflict of interest statement: No conflicts declared.

placebo-treated patients, the accuracies of predicting response and remission were similar to antidepressants. Accuracies for predicting nonresponse to placebo treatment were significantly lower than antidepressants.

Conclusions: PGMs provided clinically meaningful predictions in samples of depressed children and adolescents treated with fluoxetine or duloxetine. Future work should augment PGMs with biological data for refined predictions to guide the selection of pharmacological and psychotherapeutic treatment in children and adolescents with depression.

Keywords

Depression; adolescents; machine learning; decision support tools

Introduction

The treatment of major depressive disorder (MDD) in children and adolescents is an important public health challenge with ongoing controversies regarding the efficacy and safety of antidepressants (Walkup, 2017). Fluoxetine and escitalopram are the only FDA approved medications for MDD in adolescents, but other selective serotonin reuptake inhibitors (SSRIs) and serotonin norepinephrine reuptake inhibitors (SNRIs) are commonly prescribed off-label in clinical practice. Response to antidepressant therapy in children and adolescents is more heterogeneous when compared with adults (Chahal, Gotlib, & Guyer, 2020; Thapar, Collishaw, Pine, & Thapar, 2012). Major treatment planning challenges include determining which patients will benefit from acute antidepressant treatment, dosage increases, continuation treatment, and maintenance treatment.

One prior study of adolescents in treatment for MDD used logistic regression models to demonstrate that the rate of overall depressive symptom improvement was prognostic for acute treatment response (Tao et al., 2009). Recent work examined symptoms clusters in adolescents treated for MDD and identified two unique symptoms profiles. One profile demonstrated differential changes in depressive symptoms. A second profile failed to demonstrate differences among active and placebo treatments (Bondar, Caye, Chekroud, & Kieling, 2020). The findings from these two prior studies using The Treatment for Adolescents with Depression Study (TADS) data, albeit important, did not provide thresholds of improvement in specific symptoms to derive interpretable predictions of acute treatment course of MDD in young patients, nor did these studies achieve cross-trial replications.

Interpretable outcome predictions for children and adolescents in treatment for MDD based on symptom changes would catalyze the much-needed development of tools for shared decision making that could then be operationalized (Figure 1A) in specialty or primary care settings. These approaches are common in cardiology and other medical specialties but are underdeveloped in the child psychology and psychiatry clinical practice (Lloyd-Jones et al., 2019). This study sought to examine probabilistic graphical models (PGMs) coupled with unsupervised machine learning techniques focused on depressive symptoms to develop accurate and valid prognostic information in children and adolescents undergoing treatment for MDD. The combination of these approaches minimizes the effects of high interindividual variability (Figure 1B) in antidepressant treatment response (Athreya et al., 2021). Unsupervised learning approaches facilitate the derivation of patient stratification and symptom clusters that are sensitive to treatment effects. Further, PGMs such as hidden Markov models (HMMs) help derive compact representations of antidepressant response to generate clinically interpretable prognoses of treatment outcomes by conditioning improvements in disease severity (Athreya et al., 2021; Ceres, Schukken, & Grohn, 2020; Goyal et al., 2018; Liu, Li, Li, Song, & Rehg, 2015; Violan et al., 2020). We hypothesized that our analytical approach (Figure 1C) would (a) identify early changes in a subset of depressive symptoms that were prognostic of eventual treatment outcomes and (b) yield improved predictions of treatment outcomes across multiple classes of antidepressants in depressed children and adolescents.

Methods

Datasets

TADS (March et al., 2004) and Eli Lilly Co (Atkinson et al., 2014; Emslie et al., 2014) datasets were examined in this study (Table S1 for patient characteristics). Patients received at least 10 or 12 weeks of treatment with a study drug, fluoxetine (an SSRI), duloxetine (an SNRI), or placebo. Depressive symptoms were measured using the 17-item Children's Depression Rating Scale-Revised (CDRS-R) (Poznanski & Mokros, 1996) at baseline, 4 and 10 weeks (in Eli Lilly's datasets) or baseline 6 and 12 weeks (in TADS datasets). Early dropouts or patients in the arms of the TADS study that included cognitive behavioral therapy were not included. Each study had requisite IRB approval and clinical trial registration from respective institutions. Our current study was considered exempt by the local IRB.

PGM and prognoses rules were developed with training datasets from the TADS and Eli Lilly Co ($N = 271$) datasets (fluoxetine). The model prognostic capabilities were then examined in testing datasets from independent cohorts of patients ($N = 255$) treated for MDD with duloxetine in trials run by Eli Lilly Co (Atkinson et al., 2014; Emslie et al., 2014). Data ($N = 265$) from patients who received a pill placebo were also examined to ascertain the prognostic effects of depression symptoms that were most likely due to drug effects.

Clinical measures and outcomes

A score of ≥ 40 on the CDRS-R was the operational definition for MDD of moderate severity as this was the standard inclusion threshold for prior clinical trials enrolling adolescents with MDD. Remission was defined as a CDRS-R score ≤ 28 (Atkinson et al., 2014; Emslie et al., 2014; March et al., 2004; Mayes, Bernstein, Haley, Kennard, & Emslie, 2010; Tao et al., 2009). Response was defined as a ≥ 50% reduction (or nonresponse as a lack of ≥ 50% reduction) in CDRS-R score from baseline at 4–6 weeks or 10–12 weeks (11). If both response and remission thresholds were met, the outcome was classified as remission.

Analytical workflow

The analytical workflow (Figure 1C) to derive predictions of antidepressant response involved five steps. Steps 1–4 used training data to identify prognostic symptoms and prognoses rules. Step 5 used the prognoses rules to derive predictions in testing data.

Step 1: Construction of the PGM.—We fitted a HMM to extract the most likely variation in depression severity during the treatment for patients starting from a given baseline strata (Figure 2A). PGMs and specifically, HMMs have provided interpretable predictions of disease trajectories in prior work (Ceres et al., 2020; Goyal et al., 2018; Liu et al., 2015; Violan et al., 2020). The HMM was characterized by (a) hidden states (patient strata defined by range of total CDRS-R score), (b) observation states at 4–6 and 10–12 weeks (response status of patients in respective strata), and (c) forward transition probabilities (fraction of patients moving between strata of one time point to the next time point).

The baseline strata (hidden state) were characterized by multiple Gaussian curves. Gaussian mixture models algorithmically assigned patients to two strata with the ranges as follows, A1 (CDRS-R total score ≤ 55) and A2 (CDRS-R total score ≥ 56) (Banfield & Raftery, 1993). Clinical assessments were collected at Weeks 4–6 and Weeks 10–12 in the parent studies. Three strata were defined at treatment's intermediate or endpoints based on symptom severity thresholds (active depression, remission, and a resulting third strata of patients with scores of CDRS-R > 28 and < 40). The strata (with corresponding ranges of CDRS-R total score) at Weeks 4–6 were B1 (0–28), B2 (29–39), and B3 (≥ 40). The strata (with corresponding ranges of CDRS-R total scores) at Weeks 10–12 were C1 (0–28), C2 (29–39), and C3 (≥ 40).

All patients in B1 or C1 strata who achieved remission were also responders, 68% of patients in B2 or C2 strata were responders (without remission), and 94% of patients in B3 and C3 strata were nonresponders.

Step 2: Forward algorithm optimization for most likely depression severity trajectories.—The forward algorithm generated likelihood for all paths originating from a baseline stratum and end in a 10–12 week strata. Forward algorithm did not condition trajectories on a specific outcome of interest (e.g., remission). Instead, the construct of a forward algorithm used data from baseline and 4–6 weeks to predict the most-likely strata a patient will transition into at 10–12 weeks (see Appendix S1: Supplementary Methods for algorithm construction). For every pair of strata at baseline and treatment endpoint and all paths that connected them through a stratum at an intermediate time point, a path that had the highest likelihood and at least 10% of the patients from the baseline strata (Table S2) were chosen as the *symptom dynamic paths* (Figure 2B).

Step 3: Prognostic symptoms to predict treatment outcomes.—We defined prognostic symptoms as those that met the following three criteria: (a) non-zero symptom severities at baseline across the majority of patients (to assess the extent of early reductions in severity during treatment for predicting long-term response; Appendix S1: Supplementary Methods), (b) similar symptom severity scores (creating symptom clusters derived using hierarchical clustering for each stratum; see Figure 2C symptom clusters for A2 stratum, see Figure S1 for A1 stratum) at all time points on symptom dynamic paths originating from a baseline stratum (to establish how many symptoms with similar severity at baseline should improve at 4–6 weeks for predicting 10–12 week outcomes), and (c) different distributions of symptom severity scores between symptom dynamic paths (to quantify the level of change in a group of symptoms at 4 weeks needed to achieve specific outcomes at 10–12 weeks). These criteria facilitated the identification of depressive symptoms with similar severities across individual time points (a grouping effect) and different levels of severity between individual symptoms dynamic pathways.

Step 4: Deriving predictions of acute phase treatment outcomes using early change in severity of prognostic symptoms.—We first sought to derive thresholds of change (at 4–6 weeks) in prognostic symptoms that were needed to achieve one of the three categorical outcomes at 10–12 weeks. Using absolute difference in median scores of prognostic symptoms on symptom dynamic paths from baseline to 4- to 6-week strata, thresholds of change needed to achieve a certain outcome of interest (Table S3) was derived. We then used chi-square tests to identify the minimum number of prognostic symptoms needed to (or not) exceed thresholds at 4–6 weeks to be prognostic of outcomes at 10–12 weeks (Appendix S1: Supplementary Methods). For the prognoses rules associated with each baseline and 4–6 weeks transition, we calculated accuracy and odds ratio (OR) of the most-likely outcome expected at 10–12 weeks in Table 1. Accuracy was the fraction of patients for whom the prognoses rules predicted the correct treatment outcome. The OR represents the odds that the expected treatment outcome at 10–12 weeks would occur if patients met the prognoses rule criteria, compared with the odds of the same outcome occurring in patients not meeting the prognoses rule. Statistical significance (*p*-value) of prediction performance was derived by comparing the accuracy against the null information rate (NIR), serving as a proxy for chance. An NIR of 0.52 represents the fraction of patients in the training datasets for whom the categorical nonresponder status at 4–6 weeks correctly predicted active depression at 10–12 weeks. Finally, the Kolmogorov–Smirnov (for age) and chi-square tests (for sex and race) were used to evaluate whether prognosis rules or accuracies were associated with age, sex, or race (the common sociodemographic factors across all datasets).

Step 5: Prediction performance in testing data.—Patients in testing data were assigned to a stratum at each time point based on the mapping established in Step 1. The inputted prognostic symptom's changes at 4 weeks established the prognoses rules established in Step 5 for the prediction of outcomes at 10 weeks.

Results

We describe the results from each step of the analytical workflow.

Step 1: PGM reduces the maximum number of depression severity trajectories

By stratifying patients at each time point, we expect a maximum number of MDD response 18 paths (trajectories), that is, *N* = 2 at baseline, three at 4–6 weeks and 10–12 weeks, and *t*

= 3, and $N^t = 2*3^2 = 18$. This was a reduction in comparison with 397 unique paths without stratification (Figure 1B) – which was the motivation of Step 1 in the analytical workflow towards reducing heterogeneity of treatment outcomes.

Step 2: symptom dynamic paths (most likely depression severity trajectories)

Of the 18 possible paths in the PGM, 7 of them had the highest likelihood scores derived from the forward algorithm in Step 2 of the analytical workflow (see Figure 2B for illustration and interpretation). Patients starting in any stratum at baseline were most likely to achieve remission (i.e., be in C1 strata) at 10–12 weeks if they transitioned into the B1 stratum at 4–6 weeks, and the clinical observation at 4–6 weeks was also remission. Patients starting in the A2 strata at baseline were most likely to achieve response at 10–12 weeks, if they transitioned into the B2 or B3 stratum at 4–6 weeks and the clinical observation at 4–6 weeks was response or nonresponse respectively; and were most likely to be nonresponders at 10–12 weeks if they transitioned into the B3 stratum at 4–6 weeks and the clinical observation at 4–6 weeks was also a nonresponse. Patients starting in the A1 stratum at baseline were most likely to be nonresponders at 10–12 weeks if they transitioned into the B2 stratum at 4–6 weeks and the clinical observation at 4–6 weeks was also nonresponse (Figure 2B). Symptom cluster derived using hierarchical clustering (Figure 2C) in A2 stratum of patients with CDRS-R depression severity ≥ 56.

Step 3: Prognostic symptoms

Six CDRS-R items (difficulty having fun, social withdrawal, excessive fatigue, irritability, low self-esteem, and depressed feelings) met the prognostic symptom criteria in the training dataset. Figures 3A, B (irritability) and S2 (all other symptoms) illustrate the variations in severity of prognostic symptoms in patients with and without the superimposition of symptom dynamic paths. In patients who were not assigned to a baseline stratum (e.g., A1 or A2), there was a high degree of interpatient variation in the irritability scores, despite a mean reduction from baseline to endpoint (as shown by the large spread of boxplots in Figure 3A). By stratifying patients and then deriving symptom dynamic paths (e.g., those originating from stratum A1, as shown in Figure 3B), the discriminatory effect of scores at 10 to 12 weeks was better reflected in the patterns of response at 4–6 weeks. No such discriminatory effects occurred for nonprognostic symptoms (e.g., impaired schoolwork) or in the prognostic symptoms (inferred from fluoxetine patients) in placebo-treated patients (Figure 3C and Figure S3). No prognostic symptoms could be identified for patients who received placebo.

Step 4: Prognostic performance of prognostic symptoms in training dataset

Figure 3E and Table 1A summarize the predictive performance of the changes in prognostic symptoms at 4–6 weeks from treatment initiation.

For patients originating in the A2 stratum: (a) the accuracy in the prediction of nonresponse at 10 to 12 weeks was 62% (OR 12.3, CI 2.5–58.3, $p = .06$) by transitioning into the B3 stratum with ≥ 4 prognostic symptoms improved by ≤ 2 point at 4–6 weeks; (b) the accuracy in the prediction of response at 10–12 weeks was 86% (OR 8.4, CI 1.1–63.7, $p = 6.8\text{E}{-}05$) by transitioning into the B2 stratum with ≥ 3 prognostic symptoms improved by ≥ 2 points at 4–6 weeks; and (c) the accuracy in the prediction of remission at 10–12 weeks was 67% (OR 8, CI 0.78–81, $p = .09$) by transitioning into the B1 stratum with ≥ 4 prognostic symptoms improved by ≥ 2 points at 4–6 weeks.

For patients originating in the A1 stratum: (a) the accuracy in the prediction of nonresponse at 12/12 weeks was 66% (OR 7, CI 1.6–30.4, $p = .007$) by transitioning into the B3 stratum with ≥ 5 prognostic symptoms improved by ≤ 1 point at 4–6 weeks; (b) the accuracy in the prediction of response at 10–12 weeks was 84% (OR 5.4, CI 0.2–7, $p = .005$) by transitioning into the B2 stratum with ≥ 4 prognostic symptoms improved by ≥ 1 point at 4 weeks, and (c) the accuracy in the prediction of remission at 10–12 weeks was 72% (OR 8.3, CI 0.6–7.66, $p = .05$) by transitioning into the B1 stratum with ≥ 4 prognostic symptoms improved by ≥ 2 points at 4–6 weeks.

In over 71% (Table 1) of the patients starting from any of the baseline strata, the criteria for minimum number of prognostic symptoms needed for threshold rules were satisfied. The outcome was nonresponse for nearly all (89%) of the remaining patients.

Step 5: replication of prognostic performance of prognostic symptoms in testing datasets

Patients in the testing datasets who were treated with duloxetine were assigned to a stratum at each time point, as defined by the same range of total CDRS-R scores derived from the training dataset. The same thresholds of prognostic symptom changes at 4 weeks derived from the training cohort were applied to the testing dataset to predict outcomes at outcomes at 10 weeks (Figure 3D and 3E, additional details in Table 1B).

For patients originating in the A2 stratum: (a) the accuracies in the prediction of nonresponse at 10 weeks was 63% ($p = .02$) for patients treated with duloxetine who transitioned to the B3 stratum with ≥ 4 prognostic symptoms improved by ≤ 2 point at 4 weeks; (b) the accuracies in the prediction of response at 10 weeks was 84% ($p = 6.61E\text{-}05$) for patients who transitioned to the B2 stratum with ≥ 3 prognostic symptoms improved by ≥ 2 points at 4 weeks; and (c) the accuracies in the prediction of remission at 10 weeks was 78% ($p = .007$) for patients who transitioned to the B1 stratum with ≥ 4 prognostic symptoms improved by ≥ 2 points at 4 weeks.

For patients originating in the A1 stratum: (a) the accuracies in the prediction of nonresponse at 10 weeks was 70% ($p = .06$) for patients treated with duloxetine who transitioned to the B3 stratum with ≥ 5 prognostic symptoms improved by ≤ 1 point at 4 weeks; (b) the accuracies in the prediction of response at 10 weeks was 84% ($p = 6.6E\text{-}05$) for patients who transitioned to the B2 stratum with ≥ 4 prognostic symptoms improved by ≥ 1 points at 4 weeks; and (c) the accuracies in the prediction of remission at 10 weeks was 70% ($p = .09$) for patients who transitioned to the B1 stratum with ≥ 4 prognostic symptoms improved by ≥ 2 points at 4 weeks.

Analogous to the case in the training dataset, the prognostic symptom criteria captured variations in ≥ 77% of patients from each baseline cluster across all of the testing datasets. Nearly all (93%) of the remaining patients were nonresponders at 10 weeks. Neither age, sex, nor race was associated with chances of meeting the prognostic symptom criteria or the prediction accuracy in either fluoxetine- or duloxetine-treated patients.

Exploratory analyses: prognoses performance in placebo-treated patients

Prognostic depressive symptoms could not be identified in patients who received placebo. Table 1B summarizes the accuracy of predicting outcomes in placebo patients (assigned to baseline and 4–6 weeks strata) using the four prognostic CDRS-R-derived symptoms and compared these outcomes with fluoxetine treatment patients (Table 1A).

Discussion

This study demonstrated that PGMs can generate predictions of antidepressant response in children and adolescents with MDD across two classes of antidepressants. The PGM identified seven depression severity trajectories, wherein variations of six individual depressive symptoms (difficulty having fun, social withdrawal, excessive fatigue, irritability, low self-esteem, and depressed feelings) after 4 to 6 weeks of treatment were predictive of clinical outcomes at 10–12 weeks. The analytical approach presented in this work examined data from fluoxetine-treated patients. A replication from duloxetine-treated patients demonstrated accuracies of 77% (remission), 84% (response without remission), and 67% (nonresponse). The achieved accuracy and replication in prediction performance across antidepressant drug classes is an improvement over a prior reported accuracy of 72% (remission only) in fluoxetine-treated patients with mixed linear models (Tao et al., 2009). This preliminary work suggests that computational models have future promise for assisting clinical decisions by informing physicians on the selection, use, and dosing of antidepressants for children and adolescents with MDD.

This work represents a first step in establishing a symptom-based tool to derive interpretable predictions of treatment outcomes in children and adolescents with depression. Future efforts could integrate symptom-based PGMs with pharmacogenomic (Ramsey, Bishop, & Strawn, 2019; Ramsey et al., 2020; Troy, Poweleit, Strawn, Martin, & Ramsey, 2020), functional connectivity (Singh, Leslie, Packer, Weisman, & Gotlib, 2018), and neural metabolomic (Gabbay et al., 2012) data to enhance the predictability of treatment outcomes. In this study, placebo response did not differentiate from antidepressant response and nonresponse to placebo did not have a predictable pattern. Further improvement in symptomatology-based predictability of outcomes for placebo-treated patients would be invaluable for clinical and research efforts. For example, placebo responders could be treated with, structure, watchful waiting, and psychosocial treatments could be employed first rather than an antidepressant. Conversely, patients predicted to poorly respond to placebo could receive antidepressant treatment early. Accurate predictions of placebo response in children and adolescents with MDD could also refine clinical trial methodology (Strawn & Walkup, 2020). Protocols would then specify that all placebo responders identified by PGMs exit the trial at 4 weeks. This would result in an enriched sample to examine the true effect of the active antidepressant. Early prediction and removal of placebo responders addresses a contributor to failed pharmacotherapy trials for depression in adolescents (Walkup, 2017).

The symptom trajectories identified in the study provide additional clinical translational opportunities. Future integration of neurobiological measures into our predictive model and prospective studies offers essential opportunities to study the syndromes, phenomenology, and progression of mood disorders in adolescents. Predictive symptoms from our model (difficulty having fun, social withdrawal, excessive fatigue, irritability, low self-esteem, and depressed feelings) provide future opportunities to develop understanding of negative valence, positive valence, social processes, arousal, and regulatory systems of the Research Domain Criteria Initiative (Garvey, Avenevoli, & Anderson, 2016; Van Dam et al., 2017).

Predicting outcomes in children and adolescents treated for depression is critical in managing what could manifest into a lifelong disease burden. Recent studies raise many questions regarding the potential for over prescription, under prescription, and potential inequities of treatment for MDD in young patients (Jack et al., 2020; Sultan et al., 2018). To this end, efforts are underway to train and engage primary care physicians in the optimal treatment of MDD in children and adolescents. Algorithm-based approaches and decision support tools informed by technological advances are understudied tools that could enhance treatment approaches for children and adolescents with depression in primary care (Blanco et al., 2017; Mann, Michel, & Auerbach, 2021). Future work should extend the use of PGMs to identify prognoses rules from the Patient Health Questionnaire modified for teens (PHQ-9M) as this is the predominant rating scale used in primary care environments (Richardson et al., 2010).

The use of PGMs represents an analytical novelty in this work through the ability to derive interpretable prognoses of antidepressant treatment outcomes in depressed children and adolescents across two broad classes of antidepressants. The patient stratification at 4 to 6 or 10 to 12 weeks demonstrated consistency and relevance to clinical practice based on the distributions of treatment outcomes. Several prior efforts have used latent variable analyses with growth mixture models (Chekroud et al., 2017; Clapp et al., 2013; Gueorguieva, Mallinckrodt, & Krystal, 2011; Sakurai et al., 2013; Shelton et al., 2007; Smagula et al., 2015; Tokuoka et al., 2016; Uher et al., 2008, 2009) to derive trajectories of treatment response in depressed adults. Recent work also has used deep learning approaches to model disease trajectory (Zhang et al., 2019). In this context, mathematical constructs of PGMs have advantages as they infer most-likely trajectories of disease severity by conditioning on improvements in disease severity at intermediate time points. This approach does not require domain expertise to choose and interpret paths to ensure appropriate model fitness as in the case with growth mixture models (Bauer & Curran, 2003; Gilthorpe, Dahly, Tu, Kubzansky, & Goodman, 2014; Nylund, Asparouhov, & Muthén, 2007; Tu, Tilling, Sterne, & Gilthorpe, 2013; Wills, Silverwood, & De Stavola, 2014). The mathematical formulation presented in this work can be expanded to include extended study durations and asynchronous time points by modeling the PGM as a Markov jump process (Wang, Sontag, & Wang, 2014).

It is noteworthy that age, sex, and race did not impact prognostic symptom criteria thresholds nor prediction accuracy our model in patients treated with fluoxetine or duloxetine. This finding contrasts with prior work (Mayes et al., 2007; Strawn et al., 2021). Future research should focus on this finding to develop the current predictive model as neurobiological biomarkers are incorporated. With respect to clinical translation and implementation, a decision support tool that does not consider age, sex, nor race would have both disadvantages and advantages.

There are important limitations to consider for the interpretation of the current findings. Future work will need to explore additional categorizations of the CDRS-R at baseline and intermediate time points would enhance the predictive performance of our model. For example, more categories or different categorizations might yield superior predictive performance. Early prediction windows such as 1–2 weeks or long-term outcomes such as 6 months would have clinical utility, but this was not possible due to the structure of the training and replication datasets. As the PGM utilized data from baseline, intermediate and endpoint visits of the treatment, we did not use data from patients who dropped out in the respective clinical trials. Hence, the current findings do not account for variations in antidepressant class, dosing, patient dropouts, or variable dropout rates across studies. As such this limitation is not specific to this work but is common in all machine learning based approaches to predict outcomes (data derived after study completion) in clinical psychiatry (Chekroud et al., 2016, 2017; Iniesta et al., 2016, 2018; Koutsouleris et al., 2016). However, recent work with Bayesian hierarchical modeling suggests that antidepressant class and dose are important considerations (Suresh, Mills, Croarkin, & Strawn, 2020). The focus of this study was to establish predictability of response to pharmacotherapy for depression in child and adolescents. The methodology used in this work could however be applied to extracting response trajectories to CBT or sequential CBT combined with pharmacotherapy. Several other factors that are not considered in this work include duration of illness, adherence, family function, socioeconomic status, comorbidities, and family history of psychiatric illness. However, given the nature of the clinical trials represented in the datasets, certain assumptions are tenable. The patients were not actively using alcohol or street drugs. The children and adolescents and families were able to participate in structured follow-up appointments and assessments as dictated by the respective study protocols. Further, the patients in the current study had a valid and reliable diagnoses of MDD, did not have treatment resistant depression, and were not psychotic. We did not have data on pubertal status and could not assess its impact on prognoses performance. While the consistent prediction performances across completers of respective studies is a strength of this work, it is not possible to rule out subtle effects of adverse reactions, dropout rates, comorbidities, or additional demographic characteristics which we did not account for. Finally, our approach relies on the acceptance of exchangeability assumptions as multiple observations were pooled across clinical trial datasets (Cole & Frangakis, 2009).

Conclusion

This study provides evidence that PGMs coupled with unsupervised machine learning techniques provide clinically relevant predictive tools for children and adolescents with MDD treated with fluoxetine or duloxetine. Future efforts should examine prospectively treated patients, account for antidepressant class, examine dosing, consider psychotherapy, and integrate potential biomarkers. Finally, to further improve predictability and replicability of treatment outcomes across treatment classes beyond antidepressants, symptom-based PGMs could be integrated with pharmacogenomic, functional connectivity, and neural metabolomic data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was in part supported by the Mayo Clinic Foundation for Medical Education and Research, the National Science Foundation under award number 2041339 and the National Institute of Mental Health under award numbers R01MH113700, R01MH124655, and R01AA027486. The content is solely the responsibility of the authors and does not necessarily represent the official views of the foundation or funding agencies. The authors have declared that they have no competing or potential conflicts of interest.

References

Athreya AP, Brückl T, Binder EB, John Rush A, Biernacka J, Frye MA, … & Bobo WV (2021). Prediction of short-term antidepressant response using probabilistic graphical models with replication across multiple drugs and treatment settings. Neuropsychopharmacology, 46, 1272–1282. [PubMed: 33452433]

Atkinson SD, Prakash A, Zhang QI, Pangallo BA, Bangs ME, Emslie GJ, & March JS (2014). A double-blind efficacy and safety study of duloxetine flexible dosing in children and adolescents with major depressive disorder. Journal of Child and Adolescent Psychopharmacology, 24, 180–189. [PubMed: 24813026]

Banfield JD, & Raftery AE (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803–821.

Bauer DJ, & Curran PJ (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. Psychological Methods, 8, 338–363. [PubMed: 14596495]

Blanco C, Hoertel N, Franco S, Olfson M, He J-P, López S, … & Merikangas KR (2017). Generalizability of clinical trial results for adolescent major depressive disorder. Pediatrics, 140(6), e20161701. 10.1542/peds.2016-1701 [PubMed: 29097612]

Bondar J, Caye A, Chekroud AM, & Kieling C (2020). Symptom clusters in adolescent depression and differential response to treatment: a secondary analysis of the Treatment for Adolescents with Depression Study randomised trial. Lancet Psychiatry, 7, 337–343. [PubMed: 32199509]

Ceres KM, Schukken YH, & Grohn YT (2020). Characterizing infectious disease progression through discrete states using hidden Markov models. PLoS One, 15, e0242683. [PubMed: 33216809]

Chahal R, Gotlib IH, & Guyer AE (2020). Research Review: Brain network connectivity and the heterogeneity of depression in adolescence – A precision mental health perspective. Journal of Child Psychology and Psychiatry, 61, 1282–1298. [PubMed: 32458453]

Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, & McCarthy G (2017). Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach. JAMA Psychiatry, 74, 370–378. [PubMed: 28241180]

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, … & Corlett PR (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. Lancet Psychiatry, 3, 243–250. [PubMed: 26803397]

Clapp JD, Grubaugh AL, Allen JG, Mahoney J, Oldham JM, Fowler JC, … & Frueh BC (2013). Modeling trajectory of depressive symptoms among psychiatric inpatients: A latent growth curve approach. Journal of Clinical Psychiatry, 74, 492–499. [PubMed: 23759452]

Cole SR, & Frangakis CE (2009). The consistency statement in causal inference: A definition or an assumption? Epidemiology, 20, 3–5. [PubMed: 19234395]

Emslie GJ, Prakash A, Zhang Q, Pangallo BA, Bangs ME, & March JS (2014). A double-blind efficacy and safety study of duloxetine fixed doses in children and adolescents with major depressive disorder. Journal of Child and Adolescent Psychopharmacology, 24, 170–179. [PubMed: 24815533]

Gabbay V, Mao X, Klein RG, Ely BA, Babb JS, Panzer AM, … & Shungu DC (2012). Anterior cingulate cortex gamma-aminobutyric acid in depressed adolescents: Relationship to anhedonia. Archives of General Psychiatry, 69, 139–149. [PubMed: 21969419]

Garvey M, Avenevoli S, & Anderson K (2016). The National Institute of Mental Health Research Domain criteria and clinical research in child and adolescent psychiatry. Journal of the American Academy of Child and Adolescent Psychiatry, 55, 93–98. [PubMed: 26802775]

Gilthorpe MS, Dahly DL, Tu YK, Kubzansky LD, & Goodman E (2014). Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures. Journal of Developmental Origins of Health and Disease, 5, 197–205. [PubMed: 24901659]

Goyal D, Tjandra D, Migrino RQ, Giordani B, Syed Z, & Wiens J (2018). Characterizing heterogeneity in the progression of Alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 629–637.

Gueorguieva R, Mallinckrodt C, & Krystal JH (2011). Trajectories of depression severity in clinical trials of duloxetine: Insights into antidepressant and placebo responses. Archives of General Psychiatry, 68, 1227–1237. [PubMed: 22147842]

Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, … & Uher R (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Scientific Reports, 8, 5530. [PubMed: 29615645]

Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, … & Uher R (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. Journal of Psychiatric Research, 78, 94–102. [PubMed: 27089522]

Jack RH, Hollis C, Coupland C, Morriss R, Knaggs RD, Butler D, … & Hippisley-Cox J (2020). Incidence and prevalence of primary care antidepressant prescribing in children and young people in England, 1998–2017: A population-based cohort study. PLoS Med, 17, e1003215. [PubMed: 32697803]

Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, … & Hasan A (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry, 3, 935–946. [PubMed: 27569526]

Liu YY, Li S, Li F, Song L, & Rehg JM (2015). Efficient learning of continuous-time hidden Markov models for disease progression. Advances in Neural Information Processing Systems, 28, 3599–3607. [PubMed: 27019571]

Lloyd-Jones DM, Braun LT, Ndumele CE, Smith SC, Sperling LS, Virani SS, & Blumenthal RS (2019). Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: A special report from the American Heart Association and American College of Cardiology. Circulation, 139, e1162–e1177. [PubMed: 30586766]

Mann JJ, Michel CA, & Auerbach RP (2021). Improving suicide prevention through evidence-based strategies: A systematic review. American Journal of Psychiatry, 178(7), 611–624. [PubMed: 33596680]

March J, Silva S, Petrycki S, Curry J, Wells K, Fairbank J et al. (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents With Depression Study (TADS) randomized controlled trial. JAMA, 292, 807–820. [PubMed: 15315995]

Mayes TL, Bernstein IH, Haley CL, Kennard BD, & Emslie GJ (2010). Psychometric properties of the Children's Depression Rating Scale-Revised in adolescents. Journal of Child and Adolescent Psychopharmacology, 20, 513–516. [PubMed: 21186970]

Mayes TL, Tao R, Rintelmann JW, Carmody T, Hughes CW, Kennard BD, … & Emslie GJ (2007). Do children and adolescents have differential response rates in placebo-controlled trials of fluoxetine? CNS Spectrums, 12, 147–154. [PubMed: 17277715]

Nylund KL, Asparouhov T, & Muthén BO (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling: A Multidisciplinary Journal, 14, 535–569.

Poznanski EO, & Mokros HB (1996). (CDRS™-R) Children's Depression Rating Scale™. Revised: WPS Publish.

Ramsey LB, Bishop JR, & Strawn JR (2019). Pharmaco-genetics of treating pediatric anxiety and depression. Pharmacogenomics, 20, 867–870. [PubMed: 31453769]

Ramsey LB, Namerow LB, Bishop JR, Hicks JK, Bousman C, Croarkin PE, … & Strawn JR (2020). Thoughtful clinical use of pharmacogenetics in child and adolescent psychopharmacology. Journal of the American Academy of Child and Adolescent Psychiatry, 60, 660–664. [PubMed: 32860906]

Richardson LP, McCauley E, Grossman DC, McCarty CA, Richards J, Russo JE, … & Katon W (2010). Evaluation of the Patient Health Questionnaire-9 Item for detecting major depression among adolescents. Pediatrics, 126, 1117–1123. [PubMed: 21041282]

Sakurai H, Uchida H, Abe T, Nakajima S, Suzuki T, Pollock BG, … & Mimura M (2013). Trajectories of individual symptoms in remitters versus non-remitters with depression. Journal of Affective Disorders, 151, 506–513. [PubMed: 23886402]

Shelton RC, Prakash A, Mallinckrodt CH, Wohlreich MM, Raskin J, Robinson MJ, & Detke MJ (2007). Patterns of depressive symptom response in duloxetine-treated outpatients with mild, moderate or more severe depression. International Journal of Clinical Practice, 61, 1337–1348. [PubMed: 17627710]

Singh MK, Leslie SM, Packer MM, Weisman EF, & Gotlib IH (2018). Limbic intrinsic connectivity in depressed and high-risk youth. Journal of the American Academy of Child and Adolescent Psychiatry, 57, 775–785.e3. [PubMed: 30274652]

Smagula SF, Butters MA, Anderson SJ, Lenze EJ, Dew MA, Mulsant BH, … & Reynolds CF (2015). Antidepressant response trajectories and associated clinical prognostic factors among older adults. JAMA Psychiatry, 72, 1021–1028. [PubMed: 26288246]

Strawn JR, Mills JA, Suresh V, Peris TS, Walkup JT, & Croarkin PE (2021). Combining selective serotonin reuptake inhibitors and cognitive behavioral therapy in youth with depression and anxiety. Journal of Affective Disorders, 298(Pt A), 292–300. [PubMed: 34728290]

Strawn JR, & Walkup JT (2020). Identifying the best treatment for young people with depression. Lancet Psychiatry, 7, 562–563. [PubMed: 32563296]

Sultan RS, Correll CU, Schoenbaum M, King M, Walkup JT, & Olfson M (2018). National patterns of commonly prescribed psychotropic medications to young people. Journal of Child and Adolescent Psychopharmacology, 28, 158–165. [PubMed: 29376743]

Suresh V, Mills JA, Croarkin PE, & Strawn JR (2020). What next? A Bayesian hierarchical modeling re-examination of treatments for adolescents with selective serotonin reuptake inhibitor-resistant depression. Depress Anxiety, 37, 926–934. [PubMed: 32579280]

Tao R, Emslie G, Mayes T, Nakonezny P, Kennard B, & Hughes C (2009). Early prediction of acute antidepressant treatment response and remission in pediatric major depressive disorder. Journal of the American Academy of Child and Adolescent Psychiatry, 48, 71–78. [PubMed: 19057412]

Thapar A, Collishaw S, Pine DS, & Thapar AK (2012). Depression in adolescence. Lancet, 379, 1056–1067. [PubMed: 22305766]

Tokuoka H, Takahashi H, Ozeki A, Kuga A, Yoshikawa A, Tsuji T, & Wohlreich MM (2016). Trajectories of depression symptom improvement and associated predictor analysis: An analysis of duloxetine in double-blind placebo-controlled trials. Journal of Affective Disorders, 196, 171–180. [PubMed: 26922146]

Troy TF, Poweleit EA, Strawn JR, Martin LJ, & Ramsey LB (2020). The influence of pharmacodynamic genes on fluoxetine response in pediatric anxiety and depressive disorders. Journal of Child and Adolescent Psychopharmacology, 30, 276–277. [PubMed: 32073889]

Tu YK, Tilling K, Sterne JA, & Gilthorpe MS (2013). A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. International Journal of Epidemiology, 42, 1327–1339. [PubMed: 24038715]

Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A, … & Aitchison KJ (2008). Measuring depression: comparison and integration of three scales in the GENDEP study. Psychological Medicine, 38, 289–300. [PubMed: 17922940]

Uher R, Maier W, Hauser J, Marusic A, Schmael C, Mors O, … & McGuffin P (2009). Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. British Journal of Psychiatry, 194, 252–259.

Van Dam NT, O'Connor D, Marcelle ET, Ho EJ, Cameron Craddock R, Tobe RH, … & Milham MP (2017). Data-driven phenotypic categorization for neurobiological analyses: Beyond DSM-5 labels. Biological Psychiatry, 81, 484–494. [PubMed: 27667698]

Violán C, Fernández-Bertolín S, Guisado-Clavero M, Foguet-Boreu Q, Valderas JM, Vidal Manzano J, … & Cabrera-Bean M (2020). Five-year trajectories of multimorbidity patterns in an elderly Mediterranean population using Hidden Markov Models. Scientific Reports, 10, 16879. [PubMed: 33037233]

Walkup JT (2017). Antidepressant efficacy for depression in children and adolescents: Industry- and NIMH-funded studies. American Journal of Psychiatry, 174, 430–437. [PubMed: 28253735]

Wang X, Sontag D, & Wang F (2014). Unsupervised learning of disease progression models. Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.

Wills AK, Silverwood RJ, & De Stavola BL (2014). Comment on Tu et al 2013. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. International Journal of Epidemiology, 43, 1662–1664. [PubMed: 24585729]

Zhang XI, Chou J, Liang J, Xiao C, Zhao Y, Sarva H, … & Wang F (2019). Data-driven subtyping of Parkinson's disease using longitudinal clinical records: A cohort study. Scientific Reports, 9, 1–12. [PubMed: 30626917]

Key points

- The optimal treatment of depression in children and adolescent is a substantial public health problem.
- Machine learning and probabilistic graphical models were used to predict treatment outcomes with antidepressants in a training and testing databases.
- Variation in six depression symptoms predicted outcomes with fluoxetine or duloxetine.
- Future work should augment probabilistic graphical models with biological data to refine tools to assist decision making in clinical practice.

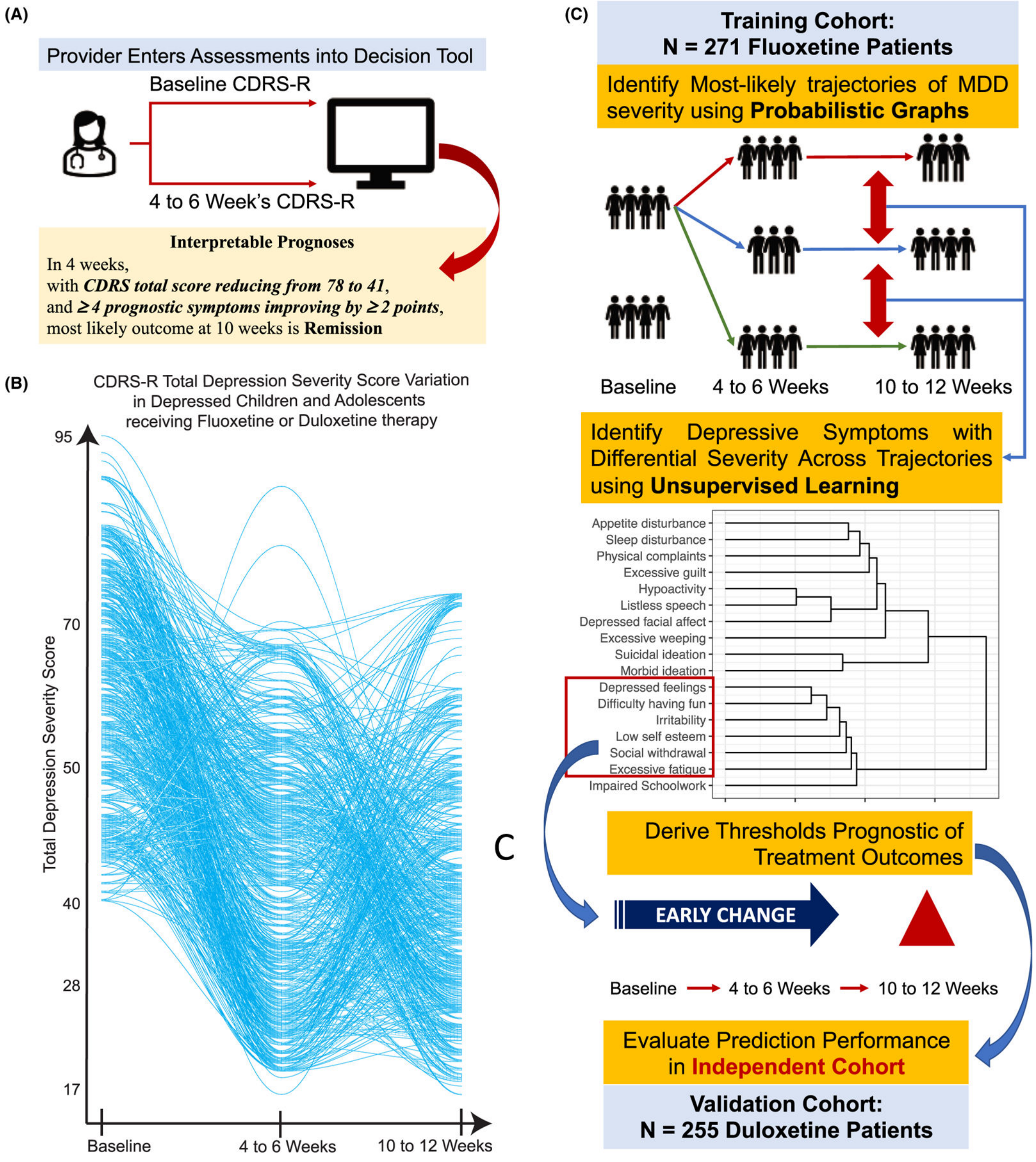

(A)
Provider Enters Assessments into Decision Tool
Baseline CDRS-R
4 to 6 Week's CDRS-R
Interpretable Prognoses
In 4 weeks,
with CDRS total score reducing from 78 to 41,
and ≥ 4 prognostic symptoms improving by ≥ 2 points,
most likely outcome at 10 weeks is Remission
(B)
CDRS-R Total Depression Severity Score Variation in Depressed Children and Adolescents receiving Fluoxetine or Duloxetine therapy
Total Depression Severity Score
95
70
50
40
28
17
Baseline
4 to 6 Weeks
10 to 12 Weeks
(C)
Training Cohort:
N = 271 Fluoxetine Patients
Identify Most-likely trajectories of MDD severity using Probabilistic Graphs
Baseline
4 to 6 Weeks
10 to 12 Weeks
Identify Depressive Symptoms with Differential Severity Across Trajectories using Unsupervised Learning
Appetite disturbance
Sleep disturbance
Physical complaints
Excessive guilt
Hypoactivity
Listless speech
Depressed facial affect
Excessive weeping
Suicidal ideation
Morbid ideation
Depressed feelings
Difficulty having fun
Irritability
Low self esteem
Social withdrawal
Excessive fatigue
Impaired Schoolwork
C
Derive Thresholds Prognostic of Treatment Outcomes
EARLY CHANGE
Baseline
4 to 6 Weeks
10 to 12 Weeks
Evaluate Prediction Performance in Independent Cohort
Validation Cohort:
N = 255 Duloxetine Patients

Figure 1.

(A) Envisioned use of proposed probabilistic graph-based tool to derive prognoses of treatment outcomes in children and adolescents treated with fluoxetine or duloxetine. (B) Trajectories of Children's Depression Rating Scale-Revised (CDRS-R) total score in patients treated with fluoxetine. (C) The machine learning workflow

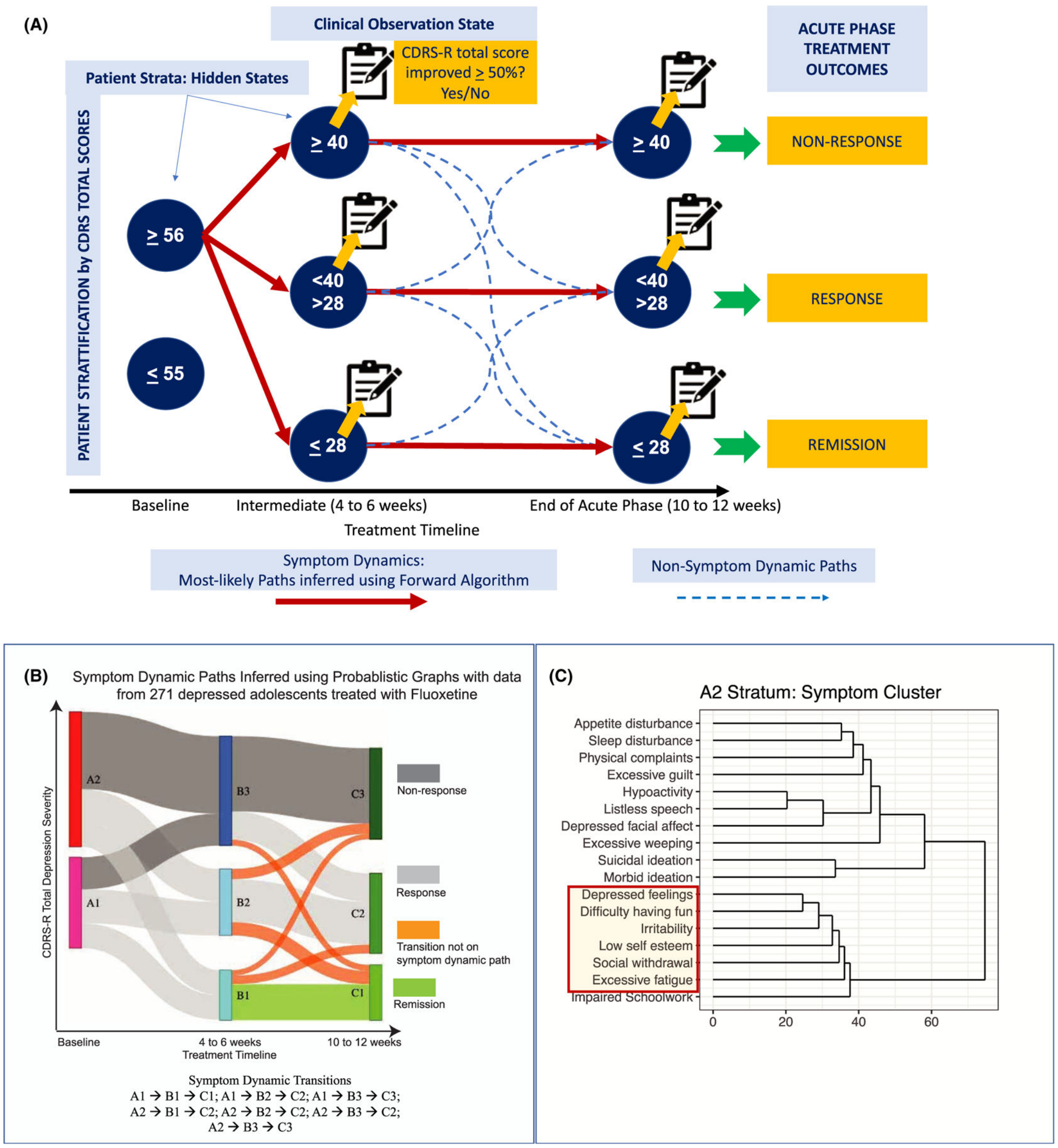
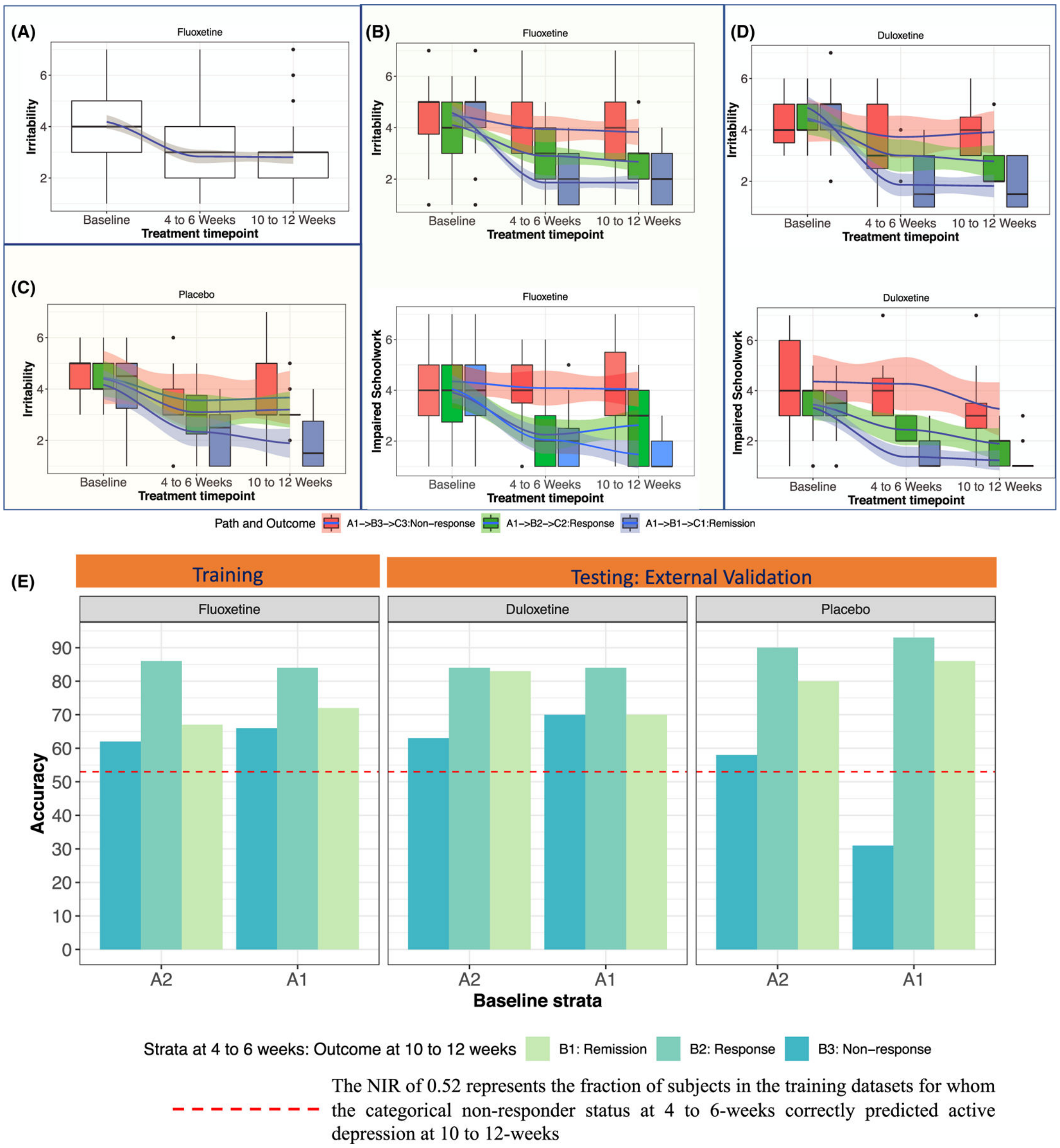

(A)
Clinical Observation State
Patient Strata: Hidden States
CDRS-R total score improved ≥ 50%? Yes/No
ACUTE PHASE TREATMENT OUTCOMES
PATIENT STRATIFICATION by CDRS TOTAL SCORES
≥ 56
≤ 55
≥ 40
<40 >28
≤ 28
NON-RESPONSE
RESPONSE
REMISSION
Baseline
Intermediate (4 to 6 weeks)
End of Acute Phase (10 to 12 weeks)
Treatment Timeline
Symptom Dynamics: Most-likely Paths inferred using Forward Algorithm
Non-Symptom Dynamic Paths
(B)
Symptom Dynamic Paths Inferred using Probablistic Graphs with data from 271 depressed adolescents treated with Fluoxetine
CDRS-R Total Depression Severity
A2
A1
B3
B2
B1
C3
C2
C1
Non-response
Response
Transition not on symptom dynamic path
Remission
Baseline
4 to 6 weeks
10 to 12 weeks
Treatment Timeline
Symptom Dynamic Transitions
A1 → B1 → C1; A1 → B2 → C2; A1 → B3 → C3;
A2 → B1 → C2; A2 → B2 → C2; A2 → B3 → C2;
A2 → B3 → C3
(C)
A2 Stratum: Symptom Cluster
Appetite disturbance
Sleep disturbance
Physical complaints
Excessive guilt
Hypoactivity
Listless speech
Depressed facial affect
Excessive weeping
Suicidal ideation
Morbid ideation
Depressed feelings
Difficulty having fun
Irritability
Low self esteem
Social withdrawal
Excessive fatigue
Impaired Schoolwork
0
20
40
60

Figure 2.

(A) Construction of the probabilistic graphical model (PGM) – a hidden Markov model comprising hidden states, observation states, and transitions. The hidden states were defined using ranges of total depression severity, wherein the ranges for depression severity at baseline were inferred using unsupervised machine learning. Observation states at the treatment's intermediate (4–6 weeks) or endpoint (10–12 weeks) record active depression [i.e., Children's Depression Rating Scale-Revised (CDRS-R) total depression severity score ≥ 40]. (B) Compact representation of CDRS-R total score variations derived using PGM. (C) Symptom clusters of patients in A2 strata based on depression severity of CDRS-R total scores less ≤ 55



The NIR of 0.52 represents the fraction of subjects in the training datasets for whom the categorical non-responder status at 4 to 6-weeks correctly predicted active depression at 10 to 12-weeks

Figure 3.

(A–D) Solid blue line was the variation in mean symptom severity, the shaded region around the solid blue line was 95% confidence interval of the variation around the mean symptom severity and the box plots visualize the variation in symptom severity at each time point. (B–D) We observe variations in symptom severity on the symptom dynamic paths originating in A1 stratum at baseline. (A) Variation in the severity of irritability in [Children's Depression Rating Scale-Revised (CDRS-R) item 4] in patients originating from A1 stratum at baseline. Although the band of the shaded region was narrow around the mean indicating reduction in symptom severity in response to therapy, the height of the boxplot and the extent of whiskers indicate high degrees of variability in scores at each time point. (B and D) Visualizing variation in prognostic (irritability) and nonprognostic symptoms (impaired schoolwork) in symptom dynamic paths of patients originating in A1 stratum and treated with either fluoxetine (Figure B) or placebo (Figure C and D). (E) Accuracy of predictions derived using short-term improvement in prognostic symptoms and CDRS-R total depression severity

Table 1

(A) Prognoses performance of prognostic symptoms in patients treated with fluoxetine making specific transitions between baseline and 4–6-week strata. (B) Prognoses performance in patients treated with duloxetine and placebo

(A) Training (TADS and Eli Lilly): Fluoxetine *N* = 271

				Prognoses rule and coverage						
Baseline strata	**Intermediate timepoint strata (4–6 weeks)**	**Number of patients making transition**	**Most-likely outcome at 10–12 weeks**	**Change in symptom severity (Baseline - 4–6 weeks)**	**Number of symptoms needing the change**	**Coverage (Fraction of patients covered by prognoses rule)**	**Probablity of most-likely outcome (Accuracy = 100*Probabilty)**	***p*-Value of accuracy with NIR = 0.52**	**Odds ratio (OR)**	**95% Confidence interval**
A2	B3	83	Non-response	≤2	≥4	0.80	.62	.06	12.30	2.5, 58.3
	B2	38	Response	≥2	≥3	0.92	.86	6.80E-05	8.40	1.1, 63.7
	B1	35	Remission	≥2	≥4	0.97	.67	.09	8.00	0.78, 81
A1	B3	46	Non-response	≤1	≥5	0.71	.66	.007	7.00	1.6, 30.4
	B2	42	Response	≥1	≥4	0.73	.84	.005	1.15	0.2, 7
	B1	27	Remission	≥2	≥4	0.77	.72	.05	8.30	1.2, 55.3

(B) Validation study

				Prognoses rule and coverage				
Baseline strata	**Intermediate timepoint strata (4–6 weeks)**	**Number of patients making transition**	**Most-likely outcome at 10–12 weeks**	**Change in symptom severity (Baseline - 4–6 weeks)**	**Number of symptoms needing the change**	**Coverage (Fraction of patients covered by prognoses rule)**	**Probablity of most-likely outcome (Accuracy = 100*Probabilty)**	***p*-value of accuracy with NIR = 0.52**
Testing: Prognoses performance at 4 weeks in Duloxetine treated patients (Eli Lilly N = 255)								
A2	B3	97	Non-response	≤2	≥4	0.85	.63	.02
	B2	38	Response	≥2	≥3	0.86	.84	6.61E-05
	B1	23	Remission	≥2	≥4	0.78	.83	.007
A1	B3	30	Non-response	≤1	≥5	0.80	.70	.06
	B2	39	Response	≥1	≥4	0.84	.84	6.60E-05
	B1	30	Remission	≥2	≥4	0.77	.70	.09
Testing: Prognoses performance at 4—6 weeks in Placebo treated Patients (Eli Lilly and TADS N = 265)								
A2	B3	119	Non-response	≤2	≥4	0.90	.58	.22
	B2	33	Response	≥2	≥3	0.93	.77	.003
	B1	18	Remission	≥2	≥4	0.88	.75	.08

(B) Validation study

Baseline strata	Intermediate timepoint strata (4–6 weeks)	Number of patients making transition	Most-likely outcome at 10–12 weeks	Prognoses rule and coverage			Probablity of most-likely outcome (Accuracy = 100*Probabilty)	*p*-value of accuracy with NIR = 0.52
				Change in symptom severity (Baseline - 4–6 weeks)	Number of symptoms needing the change	Coverage (Fraction of patients covered by prognoses rule)		
A1	B3	39	Non-response	≤1	≥5	0.62	.38	.3
	B2	33	Response	≥1	≥4	0.85	.78	.003
	B1	23	Remission	≥2	≥4	0.70	.81	.02