Learning Outcomes that Maximally Differentiate Psychiatric Treatments

Eric V. Strobl*, Semmie Kim

University of Pittsburgh

Abstract

Background: Matching each patient to the most effective treatment option(s) remains a challenging problem in psychiatry. Clinical rating scales often fail to differentiate between treatments because most treatments improve the scores of all individual items at only slightly varying degrees.

Methods: We introduce a new exploratory analysis technique called Supervised Varimax (SV). The algorithm combines the individual items that only *slightly* differ between treatments into a few scores that *greatly* differ between treatments. SV further enforces uncorrelatedness between the scores, so that they represent distinctly interpretable biopsychosocial factors. We applied SV to multi-center, double-blind, randomized and large-scale clinical trials called CATIE and STAR*D which were long thought to identify few to no differential treatment effects.

Outcomes: SV identified differential treatment effects in Phase I of CATIE ($n = 1444$, absolute sum $= 1.279$, $p = 0.0007$). Post-hoc analyses revealed that olanzapine is more effective than quetiapine and ziprasidone for hostility in chronic schizophrenia (range $= 0.284$, $p_{FDR} = 0.046$; range $= 0.283$, $p_{FDR} = 0.046$), and perphenazine is more effective than ziprasidone for emotional dysregulation (range $= 0.313$, $p_{FDR} = 0.046$). SV also discovered that buproprion augmentation is more effective than buspirone augmentation for treatment-resistant depression with increased appetite from Level 2 of STAR*D ($n = 520$, range $= 0.280$, $p_{FDR} = 0.004$).

Interpretation: SV represents a powerful methodology that enables precision psychiatry from clinical trials by optimizing the outcome measures to differentiate between treatments.

Introduction

A major goal of precision psychiatry is to match each patient to the most therapeutic treatment options(s) more precisely than the current standard of care [1]. However, most randomized clinical trials (RCTs) compare treatments to placebo rather than to other available treatments. Investigators have thus in turn designed large RCTs to compare the efficacy of existing treatments in major psychiatric illnesses. For example, the CATIE trial compared the effects of multiple antipsychotic agents in chronic schizophrenia and found superior efficacy in olanzapine [2]. Similarly, the STAR*D trial compared the effects of antidepressants and cognitive therapy in treatment-resistant depression but found little to no differential treatment effects [3, 4]. Most of the RCTs therefore revealed few differences between the treatments and did not elucidate patient subtypes that could assist in treatment selection.

Investigators have since re-analyzed datasets from the aforementioned and similar RCTs using progressively more complicated machine learning algorithms in order to reveal patient-specific differences in treatment effects. Many existing methods now utilize either a multitude of baseline clinical and/or biological variables to predict treatment response (e.g, [5, 6, 7, 8, 9, 10, 11]). Most of these studies use the remission status or a total symptom severity score as the dependent variable. Later methods achieved finer granularity by decomposing the total score into clusters

*Corresponding author

of symptoms [12, 13, 14] or considering trajectories of symptom response [15, 16, 17]. These approaches achieve superior accuracy in general by reducing noise or introducing additional dependent variables. On the other hand, the ever increasing number of variables and complexity of machine learning makes it more difficult to deploy, generalize, interpret and maintain these algorithms in everyday clinical practice. Increasing data inputs and complexity can thus also diminish real-world usability [18].

In this paper, we revisit the original problem of identifying differential treatment effects *without any independent variables other than treatment* in order to maximize pratical usability. We believe that many of the original RCTs did not identify differential treatment effects simply because the dependent variables, such as total severity scores, remission status and even validated sub-scales, were originally designed to quantify measurable symptoms rather than differentiate between treatments. As a result, treatments tend to diffusely affect nearly all items in a rating scale with weak signal and a large amount of noise. Analyzing differential treatment effects for all individual items also yields equivocal results due to the burden of multiple hypothesis testing.

We hypothesize that learning a *few* dependent variables that explicitly differentiate between treatments by amplifying relevant signals and reducing noise will lead to useful insights about treatments that can improve the standard of care without requiring additional questionnaires or measurements beyond the psychiatric interview. We specifically make the following contributions in this paper:

1. We design a new factor analysis technique called *Supervised Varimax* (SV) that takes the individual items from clinical rating scales and, unlike traditional approaches, identifies uncorrelated factors that *maximally differentiate between treatments*. We then estimate the effects of treatment on these factors, rather than on the original dependent variables.

2. We apply SV to CATIE and STAR*D in order to identify differential treatment effects on the factors. We then map the factors back onto the individual items for clinical interpretability.

3. We develop corresponding omnibus and post-hoc permutation tests that account for the learning of the optimal outcomes. We find that olanzapine is significantly more effective than quetiapine and ziprasidone for hostility in chronic schizophrenia, and perphenazine is significantly more effective than ziprasidone for emotional dysregulation in the same illness. Furthermore, buproprion augmentation is superior to buspirone augmentation in patients with treatment-resistant depression and increased appetite.

4. We finally glean simple, clinically-usable rules from the results that maximize treatment response using baseline data alone.

SV achieves high statistical power because it effectively models the heterogeneity of mental illnesses. The algorithm deconvolutes mental illness into a small number of latent factors that optimally combine all items, rather than cluster items into disjoint categories. The factors likely correspond to complex biopsychosocial processes that can only be approximately gleaned from existing clinical rating scales. Moreover, SV represents each patient as a weighted combination of these complex processes, rather than as a single category or biotype.

Methods

Clinical Trials

We analyzed a large-scale randomized clinical trial of schizophrenia called Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) [2], and another large-scale clinical trial of treatment-resistant depression called

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) [3, 4]. Both of these trials have been described in detail elsewhere. We included all patients with complete baseline data relevant to our analyses and always performed an intention-to-treat (last observation carried forward) analysis. We provide a brief summary of the components of the trials relevant to this paper below:

1. CATIE (ClinicalTrials.gov, NCT00014001, [2]) was a multi-center, double-blind, randomized clinical trial that compared atypical and typical antipsychotics in adult patients with chronic schizophrenia. We focus on Phase I of CATIE, where patients randomly received one of five treatment options: quetiapine, perphenazine, olanzapine, risperidone and ziprasidone.

2. STAR*D (ClinicalTrials.gov, NCT00021528, [3, 4]) was a multi-center, double-blind, randomized clinical trial that aimed to identify the most effective treatments for adult patients with depression whose symptoms did not remit after an initial prescription of citalopram. We focus on Level 2 of the STAR*D dataset, where participants received treatment only if they agreed to at least one of the following four options: medication switch, medication augmentation, cognitive therapy switch, and cognitive therapy augmentation. Patients then underwent randomization among the treatment options that they accepted. As a result, patients were strictly randomized only among (a) the medication switch options including buproprion, sertraline and venlafaxine, as well as (b) the medication augmentation options including buspirone augmentation and buproprion augmentation.

We downloaded the data of both studies from the National Institute of Mental Health (NIMH) Data Archive (https://nda.nih.gov/) with a limited access data use certificate.

Original Outcome Measures

The CATIE study tracked antispcyhotic response using the total score of the Positive and Negative Syndrome Scale (PANSS) [19]. The new algorithm described below takes individual items of a clinical rating scale as input. We thus input the values of all 30 individual items of the PANSS into the algorithm. On the other hand, the STAR*D trial tracked antidepressant response using the total score of the 16-item self-report Quick Inventory of Depressive Symptomatology (QIDS-SR) score [20]. We thus input all of the individual 16 items of the QIDS-SR into our algorithm.

Algorithm Overview & Empirical Testing

We describe the proposed SV algorithm at an intermediate level here. We offer an even simpler description in Results and an advanced description in the Supplementary Materials. We consider a principal components or factor analysis model, where a set of orthonormal factors $\boldsymbol{F}$ have causal effect sizes W on a set of items $\boldsymbol{Y}$ (Figure 1 (a)). Each factor is a linear combination of the items in $\boldsymbol{Y}$ rather than a cluster. However, unlike traditional PCA or factor analysis, we also consider a third layer, where binary treatments have causal effect sizes M on the factors (Figure 1 (b)). The matrices M and W are both very dense in general.

Supervised Varimax (SV) first learns the set of orthonormal factors $\boldsymbol{F}$ using principal components analysis. SV then rotates the matrix M using the rotation matrix R found by Varimax [21] to make each column of MR as sparse and as different from the other columns as possible. This ensures that each treatment then affects a small and distinct set of rotated factors $\boldsymbol{F}^*$ (Figure 1 (c)). The rotated factors are still orthonormal and still correspond to linear combinations of the items in $\boldsymbol{Y}$, but now the factors also maximally differentiate between treatments. We thus also call $\boldsymbol{F}^*$ the *optimal outcomes*.

We assessed the performance of SV by comparing it against alternative algorithms, including PCA [22], PCA with Varimax (PCA+VM) [23], factor analysis with Varimax (FA+VM) [21] and independent component analysis (ICA) [24]. We generated 1000 random factor models, as described in the Supplementary Materials, and ran the algorithms

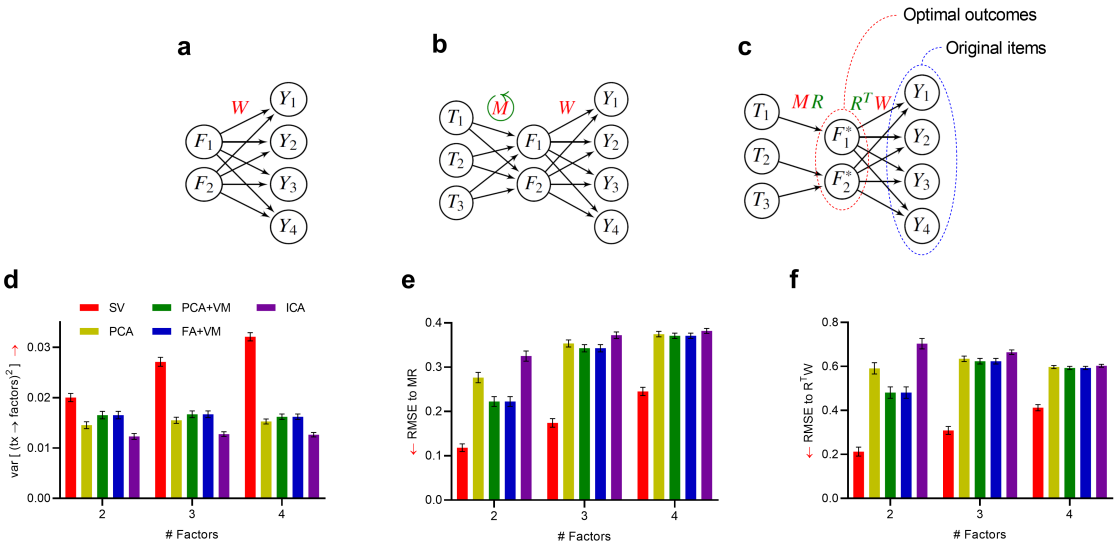

Figure 1: Algorithm overview and synthetic data results. (a) The traditional PCA or factor analysis model where factors $\boldsymbol{F}$ have causal effect sizes $\boldsymbol{W}$ on the dependent variables $\boldsymbol{Y}$. (b) SV augments the model in (a) using treatments $\boldsymbol{T}$ with causal effects $\boldsymbol{M}$ on $\boldsymbol{F}$. The algorithm then applies a varimax rotation to $\boldsymbol{M}$. (c) The rotation makes each column of the rotated causal effects $\boldsymbol{MR}$ sparse and different. As a result, only a few distinct treatments now cause each factor in $\boldsymbol{F}^*$, or the optimal outcomes. (d) SV identified the sparsest matrix $\boldsymbol{MR}$ across 1000 simulated models. SV also computed the matrix $\boldsymbol{MR}$ and the causal effects $\boldsymbol{R}^T\boldsymbol{W}$ from factors to $\boldsymbol{Y}$ with the highest accuracy in (e) and (f), respectively. Error bars denotes 95% confidence intervals of the mean. Red arrows in (d) - (f) denote direction of better performance.

on 1000 samples generated from each model. SV identified the sparsest matrix MR as assessed by the mean variance of the columns of $(MR)^{(2)}$, where we have squared each element of the matrix, and a higher such variance corresponds to increased sparsity (Figure 1 (d)). SV expectedly identified sparser matrices with more factors, whereas other methods did not display this improvement. The algorithm also estimated the matrices MR and R^TW with the lowest root mean square error (RMSE) to their ground truths (Figures 1 (e) and (f)). Each of these three comparisons held at a Bonferonni corrected threshold of 0.05/4 according to paired t-tests, since we compared SV against a total of four other algorithms. We conclude that SV outperformed all other algorithms.

Nuisance, Independent and Dependent Variables

We set the independent variables as binary treatment assignment, and the dependent variables as the individual items of the original clinical rating scales. We also set age and sex as nuisance variables and therefore partialed out these variables from each rating scale item using ordinary least squares regression before performing downstream analyses.

Hypothesis Testing

SV learns the optimal outcomes, so we need to account for the inflated Type I error rate that can result from the estimation process. We thus constructed an omnibus and a post-hoc permutation test to compute p-values that account for the estimation of optimal outcomes. The null hypothesis of both permutation tests corresponds to treatment exchangeability and therefore no differential treatment effect. The alternative hypothesis for the omnibus permutation test corresponds to the existence of a differential treatment effect across any of the tested treatments and factors. The omnibus test uses the *absolute sum* statistic corresponding to $\sum_{ij} |(MR)_{ij}|$, where i indexes the treatments, and j indexes the factors.

If we reject the omnibus null hypothesis, then we subsequently perform post-hoc permutation tests, where the alternative hypothesis corresponds to a differential treatment effect for at least one pair of medications in a particular learned factor j. The post-hoc test uses the *range* statistic corresponding to $\max_i (MR)_{ij} - \min_i (MR)_{ij}$ for factor j. We run the post-hoc test for each factor and therefore control for multiple comparisons using the false discovery rate (FDR).

We permuted the treatments and reran SV on the permuted dataset 10,000 times for both the omnibus and post-hoc tests. We provide further details of both the omnibus and post-hoc tests in the Supplementary Materials.

Clinically Significant Factors

SV learns m optimal outcomes, where m is equal to the number of treatment options. However, typically only a subset of these m factors are clinically meaningful. We thus propose a diagnostic approach to identify the clinically significant factors. We first learn the m optimal outcomes using the SV algorithm. We then plot the optimal outcomes against the variances of their columns in $(MR)^{(2)}$, since Varimax maximizes the sum of these variances (Expression (5)). We plot the variances in decreasing order. We then eliminate all factors at and below the elbow point of the resultant graph; the elbow point is often very close to zero in practice. For example, we eliminate the optimal outcomes associated with the fourth and fifth smallest variances in Figure 2 (c).

Code Availability

All R code needed to replicate the experimental results is available at github.com/ericstrobl/SV.

Role of Funding Source

No funding source provided assistance in the study design, data analysis, results interpretation, writing or submission of this report.

Results

Main Idea

We describe the SV algorithm in three levels of increasing difficulty: the simplest here, an intermediate description in the Methods, and an advanced description in the Supplementary Materials. Briefly, the *original outcome* of most clinical trials corresponds to the remission status or the total severity score according to a clinical rating scale. SV instead considers the model in Figure 1 (b), where treatments $\boldsymbol{T}$ affect items on a rating scale $\boldsymbol{Y}$ by intermediately affecting a set of latent factors $\boldsymbol{F}$ representing unknown biopsychosocial processes. In general, treatments affect the factors $\boldsymbol{F}$ in complicated ways, and the factors $\boldsymbol{F}$ affect the items $\boldsymbol{Y}$ in complicated ways. This complexity diffusely distributes the treatment effects across many factors and items which makes it difficult to differentiate between treatments. SV organizes the complexity by learning an optimal set of factors $\boldsymbol{F}^*$ from the data such that the causal effects from treatments to $\boldsymbol{F}^*$ are maximally different and simple (Figure 1 (c)). For example, treatment T_1 affects all factors $\boldsymbol{F}$ in Figure 1 (b) but only affects one optimal factor F_1^* in Figure 1 (c). The optimized factors $\boldsymbol{F}^*$, which we also call *optimal outcomes*, thus now differentiate between treatments.

Differentiating Antipsychotics

We first applied SV to Phase I of the CATIE trial, where investigators randomly assigned patients with chronic schizophrenia to five different antipsychotics including: quetiapine, perphenazine, olanzapine, risperidone and ziprasidone. Investigators then tracked the responses of patients using the PANSS score up to 18 months. 1444 patients had complete treatment assignment, age, sex and PANSS item scores at baseline in Phase I. We plot the original Phase I results in Figure 2 (a) after partialing out age and sex as nuisance variables. The CATIE trial identified olanzapine as superior to the other antipsychotics by week 18 according to the change in total PANSS score, but the trial could not differentiate between the other antipsychotics. We thus sought to identify optimal outcomes that could differentiate treatment response using the 30 individual items of the PANSS at week 18.

We first tested whether SV outperformed existing algorithms in this dataset by testing each algorithm on 1000 bootstrap datasets. We assessed how well the algorithms differentiate between treatments by computing the mean variance across the columns of $(MR)^{(2)}$ similar to Figure 1 (d), where a higher value corresponds to sparser treatment differentiation. We found that SV achieved the largest variance with all learned numbers of factors (Figure 2 (b)). Further, the performance of SV only continued to improve with an increasing number of learned factors, whereas other algorithms plateaued. We conclude that SV estimated the sparsest matrix MR in the CATIE dataset, and the results of SV mimicked those seen with the synthetic data.

Having verified the superiority of the SV algorithm, we then learned optimal outcomes that differentiate treatment response using SV and the full dataset. Diagnostics suggested three clinically significant factors based on the variance of each column of $(MR)^{(2)}$ (Figure 2 (b)). We plot the effect sizes $R^T W$ from the three factors to individual PANSS items in Figure 2 (c) for human interpretability. The first factor in red had large positive effects on items related to hostility, such as hostility itself, uncooperativeness, lack of insight and poor impulse control. The second factor in blue captured emotional dysregulation with causal effects on items related to low mood and high anxiety. Finally, the third factor in green involved negative symptoms. The three factors thus mapped onto interpretable types of dysfunction seen in schizophrenia.

We next permuted the treatments 10,000 times to test for any differential treatment effects. We rejected the omnibus null hypothesis of no differential treatment effects (absolute sum $= 1.279, p = 0.0004$). We then proceeded to test each column of MR to identify specific pairs of treatments with differential effects. We summarize the learned effect sizes in Figure 2 (e). Note that all treatments had therapeutic effects on all factors. A red cell in Figure 2 (e) does not mean a detrimental or adverse effect on the factor, but just a worse therapeutic effect than other treatments. Olanzapine had a superior effect on hostility relative to quetiapine and ziprasidone (range $= 0.284, p_{FDR} = 0.046$; range $= 0.283, p_{FDR} = 0.046$). Further, perphenazine had a superior effect on emotional dysregulation relative to ziprasidone (range $= 0.313, p_{FDR} = 0.046$). Treatments had no significant differential effects on negative symptoms. We conclude that SV identified significant differential treatment effects in two of the three factors.

We further tested whether we could translate the above results to everyday clinical practice. Recall that we learned the optimal outcomes using week 18 data, but we wanted to test whether we could predict treatment response by deriving purposefully simple, non-optimized, non-machine learning based rules with the baseline data and insight gained from SV. SV discovered that olanzapine is effective for hostility, which we coarsely scored by summing items 7, 22, 26 and 28 of the PANSS. When we give olanzapine to patients with such a hostility score above the median value at baseline, then their hostility score decreases more and faster than patients given quetiapine or ziprasidone (Figure 2 (f) red). Similarly, if we give perphenazine to patients with emotional dysregulation (sum of items 16 and 20) above the median at baseline, then they also improve more and faster than patients given ziprasidone (black). In contrast, the change in total PANSS score in patients with high hostility and high emotional dysregulation does not differ by much from Figure 2 (a) (Supplementary Materials). We conclude that the insights gained from SV on week 18

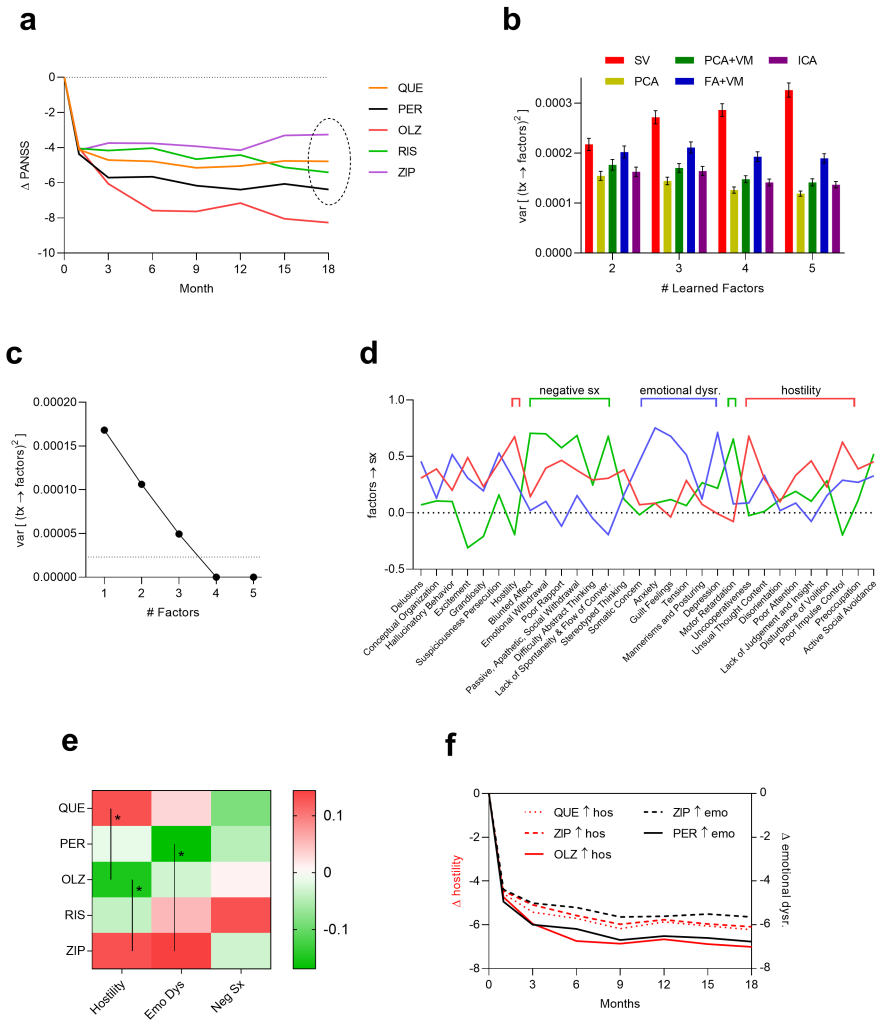

Figure 2: CATIE trial results. (a) The original analysis only distinguished olanzapine in red from all other antipsychotics using the total PANSS score. (b) SV outperformed all other algorithms in detecting differential treatment effect in this dataset regardless of the number of factors learned. (c) Diagnostics suggested the presence of three clinically significant factors after learning all five factors. (d) The clinically significant factors encapsulated hostility, emotional dysregulation and negative symptoms in chronic schizophrenia. (e) A heatmap of MR. Permutation testing revealed that olanzapine is more effective than quetiapine and ziprasidone for hostility. Moreover, perphenazine is more effective than ziprasidone for emotional dysregulation. Asterisks denote an FDR of less than 0.05. (f) Treating patients with high hostility using olanzapine resulted in a greater reduction in hostility than treating them with quetiapine or ziprasidone (red). Similarly, treating patients with high emotional dysregulation using perphenazine resulted in greater symptom reduction than ziprasidone (black).

data predict treatment response in an appropriate baseline subscore. Moreover, we can derive simple rules that match clinical common sense: giving patients treatments that best target their given constellation of symptoms improves those symptoms.

Differentiating Antidepressants

We next sought to identify differential effects of antidepressants in treatment-resistant depression using Level 2 STAR*D data, given the success of SV in identifying differential effects of antipsychotics in chronic schizophrenia using CATIE. 1312 patients had complete treatment assignment, age, sex and QIDS-SR item scores in Level 2. The STAR*D dataset is known to be exceptionally challenging, and many investigators have resorted to sophisticated machine learning algorithms in order to accurately match patients better than chance. Visual inspection of the original treatment response curves reveal why – unlike Figure 2 (a), all curves in Figure 3 (a) are near identical.

Patients in Level 2 unfortunately did not undergo strict randomization because they could accept to switch to a different medication, augment with another medication, switch to cognitive therapy, or augment with cognitive therapy (or any combination). We therefore separately analyzed only the switch medications and only the augmentation medications where strict randomization took place.

We first ran SV on the switch medications. Unfortunately, factor analysis with Varimax estimated a sparser matrix MR than SV according to the variance of the columns of $(MR)^{(2)}$ across all numbers of learned factors (Figure 3 (b)). Moreover, diagnostics with SV suggested the presence of two clinically significant factors from the three learnable factors (Figure 3 (c)), but omnibus testing failed to reject the null ($n = 659, \text{absolute sum} = 0.374, p = 0.247$). Post-hoc testing also could not differentiate between the medication switch options at even trend level with either of the two factors ($p > 0.10$ in both cases). We conclude that SV did not detect differential treatment effects among the switch medications.

SV, however, outperformed all other algorithms in estimating a sparser MR among the augmentation medications, including factor analysis with Varimax (paired t-test, $t = 12.11$, $p < 0.0001$), even with only two learned factors. Diagnostics suggested the presence of one clinically significant factor among the two learned factors (Figure 3 (c)). SV identified a differential effect between buspirone and buproprion augmentation with the one factor ($n = 520, \text{range} = 0.280, p_{FDR} = 0.004$), Figure 3 (e)). The estimated causal effects $R^T W$ corresponded to depression with increased appetite (Figure 3 (d)).

We next tested the clinical usefulness of the augmentation result by comparing the outcomes of patients with increased appetite at baseline who were also randomized to buspirone or buproprion augmentation. We specifically quantified increased appetite by the sum of items 7 and 9 in QIDS-SR. We plot the results in Figure 3 (f). Patients with increased appetite at baseline who received buproprion augmentation had larger decreases in appetite than patients who received buspirone augmentation. In contrast, the total QIDS-SR score did not consistently capture the differential effect similar to Figure 3 (a) (Supplementary Materials). We can thus recapitulate the superior effect of buproprion augmentation on appetite with a simple subscore that identifies patients with increased appetite at baseline.

Discussion

Total severity scores and remission statuses derived from clinical rating scales are not optimized to differentiate between treatments. We thus introduced the Supervised Varimax (SV) algorithm to optimize the dependent variables instead of the independent ones – unlike traditional machine learning approaches. SV transforms the individual items of a clinical rating scale into optimal outcomes that maximize differential treatment effect. The algorithm thus can detect subtle differences between the medications, even when the differences are noisily interspersed across multiple individual items. We identified differences in the treatment effects of olanzapine, perphenazine and quetiapine in

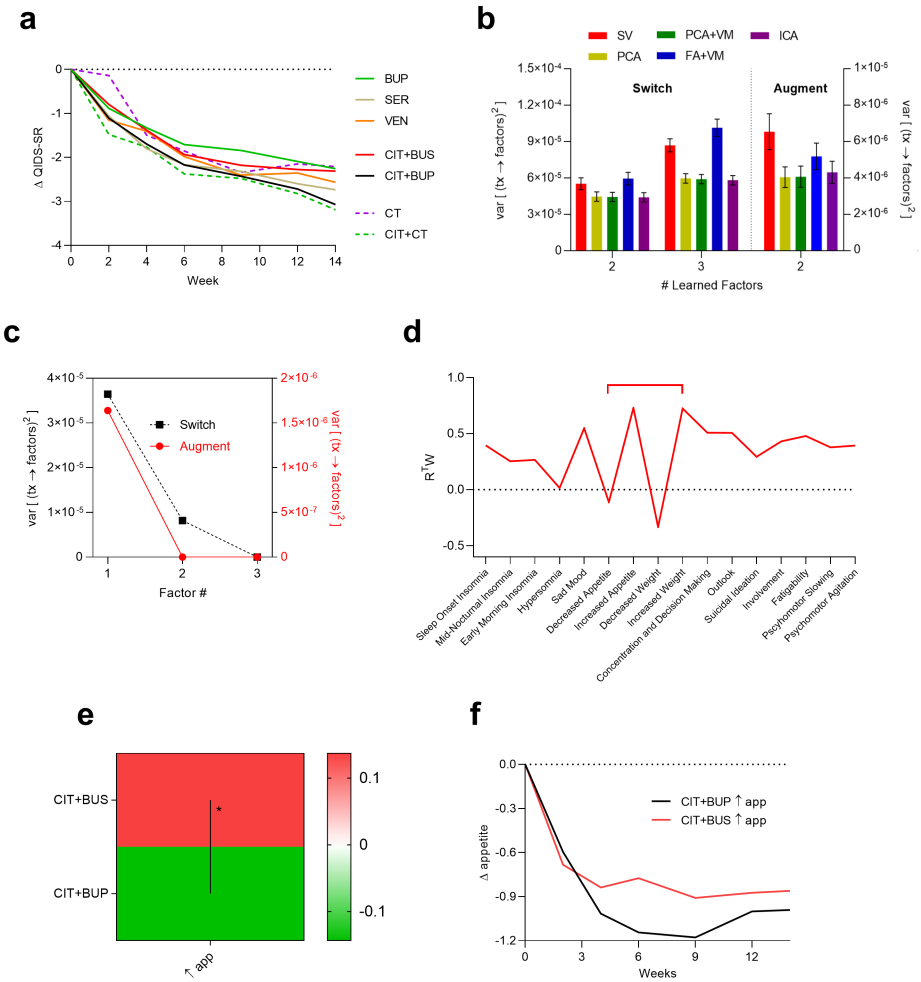

Figure 3: Level 2 STAR*D trial results. (a) All treatments had similar response curves across 18 weeks of treatment. (b) SV only outperformed the other algorithms among the augmentation medications, suggesting no detectable differential treatment effects within the switch options. (c) Diagnostics suggested the presence of two clinically significant factors and one clinically significant factor for medication switch and augmentation, respectively. (d) The one factor for augmentation corresponded to depression with increased appetite. (e) A heatmap of MR, where hypothesis testing revealed a significant differential treatment effect between buspirone and buproprion augmentation. (f) A simple clinical rule identified patients with increased appetite at baseline and recapitulated the superior efficacy of buproprion augmentation on appetite relative to buspirone augmentation.

chronic schizophrenia. We also identified buproprion augmentation as particularly effective in patients with treatment-resistant depression and increased appetite. Importantly, we detected these differential treatment effects without any independent variables other than treatment. SV thus does not require deploying, generalizing or maintaining a complex machine learning model.

Note, however, that we do not discount the importance of multiple independent variables. In fact, future work should consider learning the optimal transformation of the independent variables *and* the optimal transformation of the dependent variables in order to maximize predictive performance. We do, nevertheless, claim that *much* more emphasis has been placed on finding transformations of the independent variables rather than on learning the best outcome variables that maximize differential treatment effect. Most existing works only use fixed rating scale scores as the outcome [5, 6], or learn factors/clusters from baseline items in an unsupervised fashion [13, 25, 26]. In this paper, we showed that introducing a supervisory signal from the treatments can substantially improve the learning of outcome measures that differentiate treatments, even without any predictors other than treatment.

Other methods designed to detect differential treatment effects have mostly been restricted to major depression due to the greater availability of large clinical trial datasets in this condition [27]. However, we only found differential treatment effects in the augmentation strategies of STAR*D, in contrast to the results with CATIE. These results suggest that we may have an easier time identifying differential treatment effects in other illnesses and advanced treatment strategies. We thus encourage investigators to explore mental illnesses beyond depression and consider non-conventional treatment options more frequently despite the smaller sample sizes.

The results of SV are congruent with results from large meta-analyses, secondary analyses of adverse effects and clinical intuition. For example, SV identified olanzapine as superior to risperidone and ziprasidone for hostility in chronic schizophrenia. Two network meta-analyses have shown that olanzapine has superior efficacy over several antipsychotics in acute agitation in schizophrenia [28, 29]. SV also discovered that perphenazine is particularly effective for emotional dysregulation in chronic schizophrenia; perphenazine is one of the most well-studied first generation antipsychotics in psychotic depression [30]. Furthermore, buproprion augmentation simultaneously treats depression and reduces appetite [31]. SV identified all of these results from only two clinical trials and directly from the rating scales used in the primary analyses.

From an algorithmic standpoint, SV does not seek to maximize probabilistic independence between factors like ICA [24]. Instead, the algorithm only identifies orthogonal factors, or independence up to the second moment. We are not interested in identifying factors that satisfy all mathematical constraints associated with independence, but only enough constraints so that the factors correspond to roughly distinct biopsychosocial processes that lead to clinically actionable insights. SV thus leverages the additional rotational indeterminancy of orthogonality to identify differential treatment effects rather than maximize independence.

We must temper the above strengths with some limitations. SV imposes a linear model, even though the effect from medications to items may depend non-linearly on the dose of each medication. As a result, SV can miss complex non-linear interactions among factors. We also limited the present study to pre-specified diagnoses and associated rating scales, even though increased diversity in the dependent variables can introduce more degrees of freedom to differentiate treatment effects. Third, SV currently requires data from randomized clinical trials, even though the algorithm may benefit substantially from the diversity and large sample sizes seen in observational data with proper confounder control. Future work should therefore investigate multiple scales across multiple mental illnesses by modifying SV to perform well with observational data.

In summary, existing dependent variables are not optimized to differentiate between treatments. Most investigators have combated this issue by predicting treatment effect using many independent variables in complex machine learning models. However, we can also differentiate between treatments by simply learning dependent variables that achieve

such differentiation, such as by the SV algorithm.

References

[1] Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of 'precision psychiatry'. BMC Medicine. 2017;15:1-7.

[2] Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New England Journal of Medicine. 2005;353(12):1209-23.

[3] Trivedi MH, Fava M, Wisniewski SR, Thase ME, Quitkin F, Warden D, et al. Medication augmentation after the failure of SSRIs for depression. New England Journal of Medicine. 2006;354(12):1243-52.

[4] Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME, et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. New England Journal of Medicine. 2006;354(12):1231-42.

[5] Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry. 2016;3(3):243-50.

[6] Nie Z, Vairavan S, Narayan VA, Ye J, Li QS. Predictive modeling of treatment resistant depression using data from STAR* D and an independent clinical study. PloS One. 2018;13(6):e0197268.

[7] Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. Journal of Psychiatric Research. 2016;78:94-102.

[8] Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Scientific Reports. 2018;8(1):1-9.

[9] Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. The Lancet Psychiatry. 2016;3(10):935-46.

[10] Leighton SP, Upthegrove R, Krishnadas R, Benros ME, Broome MR, Gkoutos GV, et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. The Lancet Digital Health. 2019;1(6):e261-70.

[11] Jaworska N, De la Salle S, Ibrahim MH, Blier P, Knott V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Frontiers in psychiatry. 2019;9:768.

[12] Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine. 2017;23(1):28-38.

[13] Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA psychiatry. 2017;74(4):370-8.

[14] Kautzky A, Möller HJ, Dold M, Bartova L, Seemüller F, Laux G, et al. Combining machine learning algorithms for prediction of antidepressant treatment response. Acta Psychiatrica Scandinavica. 2021;143(1):36-49.

[15] Kok RM, van Baarsen C, Nolen WA, Heeren TJ. Early response as predictor of final remission in elderly depressed patients. International Journal of Geriatric Psychiatry. 2009;24(11):1299-303.

[16] Gueorguieva R, Chekroud AM, Krystal JH. Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: an individual patient-level data meta-analysis. The Lancet Psychiatry. 2017;4(3):230-7.

[17] Paul R, Andlauer TF, Czamara D, Hoehn D, Lucae S, Pütz B, et al. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Translational Psychiatry. 2019;9(1):187.

[18] Falkai P, Koutsouleris N. Why is it so difficult to implement precision psychiatry into clinical care? The Lancet Regional Health–Europe. 2024;43.

[19] Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia Bulletin. 1987;13(2):261-76.

[20] Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biological Psychiatry. 2003;54(5):573-83.

[21] Kaiser HF. The varimax criterion for analytic rotation in factor analysis. Psychometrika. 1958;23(3):187-200.

[22] Pearson K. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559-72.

[23] Rohe K, Zeng M. Vintage factor analysis with Varimax performs statistical inference. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2023 07;85(4):1037-60.

[24] Lee TW, Lee TW. Independent component analysis. Springer; 1998.

[25] Collins KA, Eng GK, Tural Ü, Irvin MK, Iosifescu DV, Stern ER. Affective and somatic symptom clusters in depression and their relationship to treatment outcomes in the STAR*D sample. Journal of Affective Disorders. 2022;300:469-73.

[26] Silverstein B, Patel P. Poor response to antidepressant medication of patients with depression accompanied by somatic symptomatology in the STAR* D Study. Psychiatry Research. 2011;187(1-2):121-4.

[27] Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry. 2021;20(2):154-70.

[28] Huhn M, Nikolakopoulou A, Schneider-Thoma J, Krause M, Samara M, Peter N, et al. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. The Lancet. 2019;394(10202):939-51.

[29] Paris G, Bighelli I, Deste G, Siafis S, Schneider-Thoma J, Zhu Y, et al. Short-acting intramuscular second-generation antipsychotic drugs for acutely agitated patients with schizophrenia spectrum disorders. A systematic review and network meta-analysis. Schizophrenia Research. 2021;229:3-11.

[30] Oliva V, Possidente C, De Prisco M, Fico G, Anmella G, Hidalgo-Mazzei D, et al. Pharmacological treatments for psychotic depression: a systematic review and network meta-analysis. The Lancet Psychiatry. 2024;11(3):210-20.

[31] Mohamed S, Johnson GR, Chen P, Hicks PB, Davis LL, Yoon J, et al. Effect of antidepressant switching vs augmentation on remission among patients with major depressive disorder unresponsive to antidepressant treatment: the VAST-D randomized clinical trial. JAMA. 2017;318(2):132-45.

[32] Thurstone LL. Multiple-factor Analysis: A Development and Expansion of The Vectors of the Mind. University of Chicago Committee on Publications in Biology and Medicine. Publications. University of Chicago Press; 1947.

[33] Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(3):479-98.

Supplementary Materials

Model

We use italicized letters like A to denote a single random variable, and bold italicized letters like $\boldsymbol{A}$ to denote a set of random variables. We consider a supervised factor analysis model, where m treatment assignments $\boldsymbol{T}$ causally affect q orthonormal factors $\boldsymbol{F}$ that in turn cause the p dependent variables $\boldsymbol{Y}$ (Figure 1 (b)). We require $q \leq m$ and $q \leq p$. The treatments causally affect the dependent variables $\boldsymbol{Y}$ as represented by the following structural equation:

$$\boldsymbol{Y} = \boldsymbol{T}\beta + \boldsymbol{E_Y},$$

where $\boldsymbol{E_Y}$ denotes a vector of independent and identically distributed (i.i.d.) error terms with mean zero and covariance Σ. The error terms do not necessarily follow a Gaussian distribution, and Σ may have non-zero off-diagonal elements.

The outcome $\boldsymbol{Y}$ may contain many correlated variables, which we can transform into a set of orthogonal ones by, for example, principal component analysis of $\boldsymbol{Y}$:

$$\boldsymbol{F} = \boldsymbol{Y}V\Lambda^{-1/2} = \boldsymbol{T}\beta V\Lambda^{-1/2} + \boldsymbol{E_Y}V\Lambda^{-1/2}, \tag{1}$$

where V is a $p \times q$ matrix of q eigenvectors, and Λ denotes a diagonal matrix of q non-negative eigenvalues. The set $\boldsymbol{F}$ then contains (unrotated) orthonormal factors. If we multiply $\boldsymbol{F}$ by $\Lambda^{1/2}V^T$, then we obtain:

$$\boldsymbol{F}\Lambda^{1/2}V^T = \left(\boldsymbol{Y}V\Lambda^{-1/2}\right)\Lambda^{1/2}V^T = \boldsymbol{Y}VV^T = \left(USV^T\right)VV^T = USV^T = \boldsymbol{Y},$$

where the third equality follows by the singular value decomposition $\boldsymbol{Y} = USV^T$. We thus have $\boldsymbol{Y}VV^T = \boldsymbol{Y}$. Now let $M = \beta V\Lambda^{-1/2}$ and $W = \Lambda^{1/2}V^T$, so that we recover the following model depicted in Figure 1 (b):

$$\boldsymbol{Y} = \boldsymbol{Y}VV^T = \boldsymbol{T}MW + \boldsymbol{E_Y}, \tag{2}$$

since $\boldsymbol{E_Y}VV^T$ has covariance matrix $VV^T\Sigma VV^T = VV^T\left(V\Lambda V^T\right)VV^T = V\Lambda V^T = \Sigma$.

Optimal Rotation

The transformation matrix $V\Lambda^{-1/2}$ in Equation (1) is not unique because the columns of $\boldsymbol{F}R$ are orthonormal as well, where R corresponds to an orthonormal rotation matrix. We thus also consider any transformation matrix $V\Lambda^{-1/2}R$:

$$\boldsymbol{F}^* = \boldsymbol{F}R = \boldsymbol{T}MR + \boldsymbol{E_Y}V\Lambda^{-1/2}R, \tag{3}$$

where $\boldsymbol{F}^*$ corresponds to the optimized outcomes, and $\boldsymbol{E_Y} V\Lambda^{-1/2}R$ again denotes a vector of i.i.d. error terms with mean zero but covariance $R^T\Lambda^{-1/2}V^T\Sigma V\Lambda^{-1/2}R$.

We now specify the rotation R. We let $\mathcal{R}(q)$ denote the set of $q \times q$ rotation matrices. We seek a rotation that maximally differentiates treatment effects on the set of latent factors so that:

$$\underset{R\in\mathcal{R}(q)}{\arg\min} \sum_{j<k} \left[\frac{1}{q}\sum_i (MR)_{ij}^2 (MR)_{ik}^2 - \left(\frac{1}{q}\sum_i (MR)_{ij}^2\right)\left(\frac{1}{q}\sum_i (MR)_{ik}^2\right)\right]. \tag{4}$$

In other words, we minimize the covariance of pairs of columns of treatment effects squared, and then sum over all possible pairs. As a result, each factor tends to be caused by a different set of treatments.

The minimization problem in Expression (4) yields the same solution as the following maximization problem, also known as the varimax rotation [21]:

$$\underset{R\in\mathcal{R}(q)}{\arg\max} \sum_{j} \left[\frac{1}{q}\sum_i (MR)_{ij}^4 - \left(\frac{1}{q}\sum_i (MR)_{ij}^2\right)^2\right], \tag{5}$$

where we maximize the variance of each column of $(MR)^{(2)}$ (element-wise squared). In other words, we maximize a quantity similar but not equivalent to the kurtosis of each column of MR to induce outliers and sparsity. We combine Expressions (4) and (5) to see the equivalence between the maximization and minimization as follows:

$$\begin{aligned}
(4) + 2(5) &= \sum_{j,k} \left[\frac{1}{q}\sum_i (MR)_{ij}^2 (MR)_{ik}^2 - \left(\frac{1}{q}\sum_i (MR)_{ij}^2\right)\left(\frac{1}{q}\sum_i (MR)_{ik}^2\right)\right] \\
&= \frac{1}{q}\sum_i \left(\sum_j (MR)_{ij}^2 \sum_k (MR)_{ik}^2\right) - \left(\frac{1}{q}\sum_i\sum_j (MR)_{ij}^2\right)\left(\frac{1}{q}\sum_i\sum_k (MR)_{ik}^2\right) \\
&= C,
\end{aligned}$$

which is equal to a constant C because rotations preserve row vector lengths, or the sum of squares of its elements. Hence, the minimization problem in Expression (4) yields the same solution as the maximization problem in Expression (5).

Varimax is known to approximately induce part of Thurstone's *simple structure* [32] in MR, which we paraphrase for the present context below:

1. each factor has no causal effect from most treatments;
2. each factor has large causal effects (in magnitude) from a small number of treatments;
3. few factors have large causal effects from the same treatment.

In particular, the row sums of the squared elements of MR remain fixed under rotation because R is orthonormal. We thus maximize the variance depicted in Expression (5), or the mean of squared pairwise distances, so that most squared elements of MR are large or zero; this in turn satisfies the first two items listed above. The third follows by equivalently minimizing the covariance shown in Expression (4).

Supervised Varimax

We are now ready to describe the proposed algorithm, which we summarize in Algorithm 1. SV first performs an eigendecomposition of the correlation matrix of $\boldsymbol{Y}$ and extracts the eigenvectors associated with the top m largest

eigenvalues in Line 1 so that $q = m$. We now have $\boldsymbol{F} = \boldsymbol{Y}V\Lambda^{-1/2}$ corresponding to the unrotated factors. SV then regresses $\boldsymbol{F}$ on $\boldsymbol{T}$ to obtain the causal effects M in Line 2. Next, SV sparsifies M with a varimax rotation in order to compute the optimal outcomes $\boldsymbol{F}^*$ in Lines 3 and 4, respectively. Note that Varimax has permutation and sign indeterminancies [23], which we determine in Line 5 by sorting $\boldsymbol{F}^*$ according to variance explained and non-negatively correlating $\boldsymbol{F}^*$ to $\sum_k Y_k$ via sign flips. SV thus ultimately outputs the desired matrices MR, $R^T W$ and $\boldsymbol{F}^*$ with permutation and sign determinancy.

Algorithm 1 Supervised Varimax

Input: individual items $\boldsymbol{Y}$, treatment assignment $\boldsymbol{T}$
Output: MRD, $D^T R^T W$, $\boldsymbol{F}^* D$

1: $\boldsymbol{F}, V, \Lambda \leftarrow$ eigendecomposition of the correlation matrix of $\boldsymbol{Y}$
2: $M \leftarrow$ regress $\boldsymbol{F}$ on binary treatment assignment $\boldsymbol{T}$
3: $R \leftarrow$ perform a varimax rotation on M
4: $\boldsymbol{F}^* \leftarrow \boldsymbol{F}R$
5: $D \leftarrow$ signed diagonal matrix so that (a) $\boldsymbol{F}^* D$ is sorted in decreasing order by proportion of variance of $\sum_k Y_k$ explained and (b) each entry of $\boldsymbol{F}^* D$ has a non-negative correlation with $\sum_k Y_k$

Permutation Tests

We consider the following omnibus null and alternative hypotheses written in plain English:

- $\mathcal{H}_0$: treatments are exchangeable so that no differential treatment effect exists;
- $\mathcal{H}_1$: a differential treatment effect exists for some factor.

We operationalize the above omnibus hypotheses as follows:

- $\mathcal{H}_0 : \boldsymbol{Y} \perp\!\!\!\perp T$,
- $\mathcal{H}_1 : \sum_{ij} |(MR)_{ij}| > \left(\sum_{ij} |(MR)_{ij}|\right)_{\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{T}}$.

We call $\sum_{ij} |(MR)_{ij}|$ the *absolute sum*, and the notation $\left(\sum_{ij} |(MR)_{ij}|\right)_{\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{T}}$ refers to the absolute sum when the null hypothesis holds. The permutation test then proceeds by permuting treatment assignment, computing the absolute sum on each permuted sample and then counting the proportion of cases where the statistic falls at or above the same quantity computed on the original samples.

If we reject the omnibus null hypothesis, then we also test pairs of treatments within each factor using the following post-hoc hypothesis:

- $\mathcal{H}_0$: all treatments are exchangeable so that no differential treatment effect exists;
- $\mathcal{H}_1$: a differential treatment effect exists in factor j for some pair of treatments.

We operationalize the post-hoc hypotheses as:

- $\mathcal{H}_0 : \boldsymbol{Y} \perp\!\!\!\perp T$,
- $\mathcal{H}_1 : \left(\max_i (MR)_{ij} - \min_i (MR)_{ij}\right) > \left(\max_i (MR)_{ij} - \min_i (MR)_{ij}\right)_{\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{T}}$.

We call $\max_i (MR)_{ij} - \min_i (MR)_{ij}$ the *range* statistic. The permutation test again proceeds by permuting treatment assignment, estimating MR for each permuted sample and counting the proportion of cases where the range falls at or above the same quantity computed on the original samples.

We controlled for multiple comparisons among the post-hoc tests using the false discovery rate. We converted all permutation p-values to q-values, which we denote by p_{FDR} in this paper [33]. Computing the q-values requires an estimate of the upper bound on the proportion of null hypotheses. We assume that all clinically insignificant factors satisfy the null hypothesis, so that we can only reject the null with clinically significant factors. We also only perform post-hoc testing after rejecting the omnibus null hypothesis. We thus bound the proportion of null hypotheses with $(d - r - 1)/d$, where $d \geq m$ denotes the total number of post-hoc tests performed, and r the number of clinically insignificant factors determined from diagnostics.

Synthetic Data Generation

We drew 1000 samples from the model shown in Equation (2), where each entry of $\boldsymbol{E_Y}$ followed a t-distribution with three degrees of freedom; we chose this non-Gaussian distribution to ensure identifiability of the ICA solution. We sampled the matrices M and W by drawing each entry from $\text{Unif}([-1, -0.25] \cup [0.25, 1])$. We then performed a varimax rotation on the ground truth matrix M to yield the rotation matrix R. We removed sign and permutation indeterminancies using the same procedure as Line 5 in SV. We repeated the above process 1000 times for 2, 3 and 4 factors in $\boldsymbol{F}$. We thus generated a total of $3 \times 1000 = 3000$ unique datasets.

Change in Total Scores

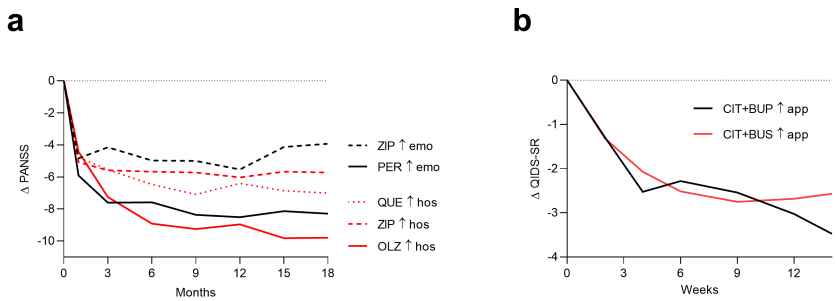

Figure 4: Total severity scores do not effectively differentiate between treatments. As a result, changes in the total score of PANSS in (a) and QIDS-SR in (b) simply mimic those seen in Figures 2 (a) and 3 (a), respectively – even in the identified subpopulations.