Estimating patient-specific treatment advantages in the 'Treatment for Adolescents with Depression Study'

Simon Foster [1, 2, 3]

Meichun Mohler-Kuo [1, 2, 3, 4]

Lynette Tay [1]

Torsten Hothorn [1]

Heidi Seibold [1]

Author affiliations

[1] Epidemiology, Biostatistics and Prevention Institute

University of Zurich

Hirschengraben 84

8001 Zürich, Switzerland

[2] Swiss Research Institute for Public Health and Addiction associated with the University of Zurich

Konradstrasse 32

8031 Zürich, Switzerland

[3] Department of Child and Adolescent Psychiatry and Psychotherapy (KJPP)

University Hospital of Psychiatry Zurich

University of Zurich

Neumünsterallee 9

8032 Zurich, Switzerland

[4] La Source, School of nursing sciences
HES-SO University of Applied Sciences and Arts of Western Switzerland
Av. Vinet 30
1004 Lausanne, Switzerland

Corresponding author

Simon Foster
Department of Child and Adolescent Psychiatry and Psychotherapy (KJPP)
University Hospital of Psychiatry Zurich
University of Zurich
Neumünsterallee 9
8032 Zurich
Phone: +41 43 556 40 04
Fax: +41 43 499 26 02
Email: simon.foster@uzh.ch

Abstract

The ‘Treatment for Adolescents with Depression Study’ (TADS, ClinicalTrials.gov, identifier: NCT00006286) was a cornerstone, randomized controlled trial evaluating the effectiveness of standard treatment options for major depression in adolescents. Whereas previous TADS analyses examined primarily effect modifications of treatment-placebo differences by various patient characteristics, less is known about the modification of inter-treatment differences, and hence, patient characteristics that might guide treatment selection. We sought to fill this gap by estimating patient-specific inter-treatment differences as a function of patients’ baseline characteristics. We did so by applying the ‘model-based random forest’, a recently-introduced machine learning-based method for evaluating effect heterogeneity that allows for the estimation of patient-specific treatment effects as a function of arbitrary baseline characteristics. Treatment conditions were cognitive-behavioural therapy (CBT) alone, fluoxetine (FLX) alone, and the combination of CBT and fluoxetine (COMB). All inter-treatment differences (CBT vs. FLX; CBT vs. COMB; FLX vs. COMB) were evaluated across 23 potential effect modifiers extracted from previous studies. Overall, FLX was superior to CBT, while COMB was superior to both CBT and FLX. Evidence for effect heterogeneity was found for the CBT-FLX difference and the FLX-COMB difference, but not for the CBT-COMB difference. Baseline depression severity modified the CBT-FLX difference; whereas baseline depression severity, patients’ treatment expectations, and childhood trauma modified the FLX-COMB difference. All modifications were quantitative rather than qualitative, however, meaning that the differences varied only in magnitude, but not direction. These findings imply that combining CBT with fluoxetine may be superior to either therapy used alone across a broad range of patients.

Keywords

adolescent; major depressive disorder; second-generation antidepressive agents; cognitive therapy; randomized controlled trial; effect heterogeneity

Introduction

The ‘Treatment for Adolescents with Depression Study’ (TADS) was a cornerstone clinical trial that evaluated the effectiveness of standard treatment options for major depression in adolescents (March et al., 2004; Thapar et al., 2012). These options included treatment with the selective serotonin reuptake inhibitor (SSRI) fluoxetine, with cognitive-behavioural therapy, and with these two treatments combined. Whereas cognitive-behavioural therapy failed to outperform placebo after 12 weeks acute-phase treatment, fluoxetine outperformed placebo and the combined therapy outperformed both the placebo and either active therapy used alone (March et al., 2004).

Whereas establishing the average effectiveness of a treatment is a necessary step, a major goal in the era of personalized and precision medicine is to identify the treatment best-suited for each individual (Simon and Perlis, 2010). Indeed, several effect-modification analyses of the TADS data have been published, assessing various patient characteristics as potential modifiers of the placebo-treatment differences (Curry et al., 2006; Kratochvil et al., 2009; Lewis et al., 2010). However, as noted by Simon and Perlis (2010), the hallmark of personalized medicine is to be able to choose appropriate treatments for each particular patient; that is, to answer the question of whether a patient with characteristic X responds better to treatment A than to treatment B. This implies effect modifications of inter-treatment differences, which were derived only indirectly and rather informally in previous studies. In addition, previous analyses were limited by the drawbacks of conventional effect-modification analyses, which particularly include testing effect modifiers individually, while disregarding all other patient characteristics; spurious effect modifications resulting from multiple hypothesis testing and *ad-hoc* stepwise model building; and the inability to estimate patient- versus subgroup-specific treatment effects (Dahabreh et al., 2016; Kent et al., 2010;

Kessler et al., 2017; Moher et al., 2012; Rothwell, 2005; Seibold et al., 2018a; Willke et al., 2012).

Seibold et al. have recently introduced a statistical approach, called the model-based random forest, to tackle effect heterogeneity (Seibold et al., 2018a). This approach blends traditional parametric modelling with a machine-learning approach, known as the random forest (hence: "model-based random forest"). Considering the perils and critiques associated with traditional approaches used in the cited studies, this new approach has a number of advantages. First, it allows for deriving a point estimate of the counter-factual treatment effect for each individual patient rather than for subgroups of patients. Second, the new approach allows for carrying out a global test for the presence of statistically significant heterogeneity within a treatment effect across the set of potential effect modifiers. Third, potential effect modifiers are considered simultaneously, rather than individually. In this way, complicated effect modifications that involve more than one effect modifier can be addressed. Failing to incorporate such multidimensional modifications has been found to be a major drawback of analyses that only consider one effect modifier at a time (Kent et al., 2010). Fourth, effect modification can be dealt with bottom-up, with no need to explicitly define subgroups or statistical interaction effects and their functional forms in advance. Finally, the spurious effect modification introduced by multiple hypothesis testing and *ad-hoc* stepwise model building is reduced by considering all effect modifier candidates simultaneously and by the set-up of the algorithm (Seibold et al., 2018a; Zeileis et al., 2008).

Our aim was to address the afore-mentioned gaps in understanding by analysing for heterogeneity in inter-treatment differences for the three active TADS treatment arms, using this novel, more comprehensive approach introduced by Seibold et al. In particular, we sought to derive point estimates for the counter-factual inter-treatment differences of patients as a function of their baseline characteristics. Our analysis thereby tested for patient characteristics that might guide choices between multiple treatments. Failing to conduct such an analysis would result in continuing to treat individuals based on estimated average treatment effects, be they average effects across all trial subjects, or average effects among specific subject subgroups.

Methods

We re-analyzed the data of the acute phase of the TADS trial (ClinicalTrials.gov, NCT00006286), which encompassed the first 12 weeks of treatment. TADS is a randomized controlled trial (RCT) that was designed to compare the effectiveness of common treatments against a pill placebo for the treatment of adolescents with a major depressive disorder (MDD). A detailed description of the trial design has already been published elsewhere (TADS Team, 2003). TADS was monitored quarterly by the data safety and monitoring board of the National Institute of Mental Health (NIMH) of the United States of America (USA). The study protocol was approved and monitored by the institutional review boards at each study site. Informed written consent was obtained from all patients and at least one of their parents. Data were acquired from the NIMH Data Archive (NDA) through a limited-access data certificate by the first author.

Study population

A sample of 439 adolescents who met the criteria for major depressive disorder (MDD) as defined in the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV, American Psychiatric Association, 1994), were recruited at 13 study sites. Patients were 12-17 years old, and 54.4% were female. Roughly 74% were white, 12.5% African American, 8.9% Hispanic, and 4.8% other. Roughly 61% reported family incomes of $40,000 or higher over the preceding 12 months. Almost 48% had at least one coexisting DMS-IV disorder. A comparison of the TADS sample against other clinical and epidemiological samples as well as detailed descriptions concerning eligibility criteria, subject recruitment, data collection, sample size calculation, and sample characteristics have been published elsewhere (March et al., 2004; TADS Team, 2003, 2005).

Interventions

Detailed descriptions of the intervention arms have been published elsewhere (March et al., 2004; TADS Team, 2003). The study's acute phase included a pill placebo (PBO); cognitive-behavioural therapy (CBT); fluoxetine (FLX); and the combination of CBT and FLX (COMB). Eligible patients were randomly assigned to one of the treatment arms, with a 1:1:1:1 allocation ratio (PBO: n=112; CBT: n=111; FLX: n=109; COMB: n=107), and with study site and sex as stratification variables (March et al., 2004). Participants and therapists remained blinded to group allocation in the FLX and PBO groups, but were aware that patients received active medication in the COMB group and no medication in the CBT group (March et al., 2004).

Outcome

Independent evaluators blinded to treatment arm assignments assessed MDD at baseline and at week 12, using the Children's Depression Rating Scale-Revised (CDRS-R). The CDRS-R is a clinician-administered, validated rating scale based on the synthesis of information collected by interviewing both the adolescent and a parent (Mayes et al., 2010; Poznanski and Mokros, 1995). The raw summary score served as a measure of depression (Curry et al., 2006; March et al., 2004).

Missing values at the end of the acute phase were replaced by scores predicted by the random coefficient regression model used in the original TADS analyses (Curry et al., 2006). This imputation method was suggested to provide a less-biased estimate of treatment outcome than the last carried forward approach (Curry et al., 2006). Missing values were replaced in sixty-one (13.9%) of the 439 patients.

Effect modifiers

We included all the patient characteristics selected by Curry et al. (2006) based on a literature review that they conducted prior to their effect modification analysis, with three adaptations. First, we replaced 'comorbidity with disruptive behaviour' with 'comorbidity with an attention-deficit/hyperactivity disorder (ADHD)', since ADHD, but not 'comorbidity with disruptive behaviour' was found to be an effect modifier (Curry et al., 2006; Kratochvil et al., 2009). Second, since childhood trauma was found to be an effect modifier in a separate TADS analysis, it also was included (Lewis et al., 2010). Finally, we also included study site. With these adaptations, a final set of 23 potential effect modifiers was included in the analyses (Table 1; see references Curry et al., 2006; TADS Team, 2005, for details). Missing values for the effect modifiers were handled by the statistical method described below.

Statistical analysis

Our analysis was an application of the recently-developed 'model-based random forest' . Details and the mathematical derivation of the method can be found in Seibold et al. (2018a). The model-based random forest is an extension of the model-based recursive partitioning algorithm outlined by Zeileis and colleagues (Seibold et al., 2016; Zeileis et al., 2008). This algorithm was recently shown to be competitive identifying effect modifications (Alemayehu et al., 2018). Succinctly, it works as follows:

1. Fit a parametric model to the data, consisting of the outcome modelled as a function of the intercept and the treatment effect (as would be done in the conventional statistical analysis of an RCT)
2. Test for instability in the model parameters across the potential effect modifiers
3. If some overall instability is detected, split the data on the effect modifier associated with the highest parameter instability; otherwise stop.
4. Within each resulting data subset (patient subgroup), repeat the algorithm.

Within each node (subgroup) of the tree, a parameter stability test is carried out for each effect modifier. If any of the tests is statistically significant, a split is carried out on the effect modifier producing the smallest p-value. Across nodes, significance is controlled by closed testing, as the tests are nested recursively, meaning that the significance of p-values is only interpretable if parameter stability has been rejected for all previous nodes. This algorithm set-up assures that only informative effect modifiers are selected, while spurious effect modification is dampened, since splits in irrelevant effect modifiers are selected in each node only with probability α, and the algorithm stops growing the tree in nodes wherever parameter stability cannot be rejected (Seibold et al., 2018a; Zeileis et al., 2008). In addition, the algorithm automatically handles missing values amongst the effect modifiers. Once a split has been implemented in an effect modifier containing missing values, the observations with missing values are randomly assigned to a daughter node following the node distribution. During further testing for parameter instability, observations with missing values are ignored.

By applying the algorithm recursively, subgroups with increasingly-homogeneous model parameters are derived. Ultimately, rather than having a single global model, several subgroup-specific models are estimated, resulting in subgroup-specific treatment effects. Crucially, because the effect modifier variables are used to define the subgroups, the model parameters of the global model can be written as a function of subgroup-defining variables. The resulting function of the treatment effect can then be understood as an estimate of the counterfactual treatment effect of a patient, with particular values on the effect modifiers (Seibold et al., 2018a).

The above-described tree-based algorithm is then extended to a random forest. In a random forest, rather than estimating a single tree, an ensemble of trees is estimated (Efron and Hastie, 2016). Each model-based tree in the forest is estimated on a randomly-drawn subsample of the original data, and each split in the tree is based on a randomly-sampled subset of candidate effect modifiers (Efron and Hastie, 2016; Seibold et al., 2018a). Random forests have the advantages of allowing for estimating smooth relationship forms, rather than being restricted to the step-functions provided by individual trees (Seibold et al., 2018a). The treatment effect can, thereby, be estimated as an arbitrary smooth function of the effect modifiers. This ability sets this approach apart from other recently-introduced tree-based and/or interaction models (Seibold et al., 2018a).

Whereas a conventional random forest would provide information on the similarity of patients with respect to the outcome variable, the model-based random forest provides information on the similarity of patients with respect to the model parameters and, hence, the treatment effect (Seibold et al., 2018a). Thus, the model-based random forest allows for measuring how similar patients are, with respect to treatment effect. Based upon this similarity, a personalized treatment effect for each patient is estimated by re-calculating the global model, with the patients in the data set weighted according to their similarity to the patient under question.

The model-based random forest provides several results (Seibold et al., 2018a). First, since patient-specific treatment effects of the sampled patients are estimated, the overall distribution of effects can be examined. A hypothesis test is provided that evaluates whether the patient-specific effects improve the global model, thereby globally testing for the presence of effect heterogeneity across effect modifiers. Second, if evidence of effect heterogeneity is present, variable importance measures can be derived that reveal each effect modifier's contribution in the random forest. Third, partial dependence plots can be derived that display patient-specific treatment effects as a function of relevant effect modifiers. Finally, the random forest can be used to predict the counterfactual treatment effects in future patients. In an RCT with several treatment arms, this includes estimating a patient's treatment effect if he or she had been allocated to another treatment arm.

As a global model, we fitted a Gaussian generalized linear model with a log-link and the log of baseline depression severity as offset (Seibold et al., 2018a). This amounts to modeling the patients' proportional change from baseline to the end of week 12. A negative treatment coefficient means that the expected change from the treatment was more favorable than that with the reference condition. Exponentiation of the model intercept produces the expected change in the reference condition; whereas exponentiation of the treatment coefficient produces the factor by which the change in the treatment condition is higher (or lower) than the change in the reference condition. Note that other parametric models could be used as global models, depending upon the type of outcome one wishes to study (Seibold et al., 2016; Seibold et al., 2018a; Zeileis et al., 2008).

A separate model-based random forest was estimated for each treatment-treatment pair. The p-values of the corresponding global effects were adjusted for multiple testing by Holm's procedure (Shaffer, 1995), as were the global tests for the presence of effect heterogeneity. An intention-to-treat sample was used in all analyses. Analyses were carried out using R software (R Core Team, 2016), based on R-code provided by Seibold et al. (2018a). This code has now been integrated into the newly released and publicly available R-package “model4you” (Seibold et al., 2018b).

Results

Overall variability of patient-specific inter-treatment differences

Table 2 displays the parameter estimates of the global models of all inter-treatment differences, summarizes the variability of the estimates, and provides p-values for the tests of overall effect heterogeneity. Considering the CBT-FLX difference, the global model indicated greater effectiveness of FLX (*b* = -0.13, 95% CI: -0.22 to -0.05, p = 0.0040). The personalized effects ranged from -0.16 to -0.11, with statistical evidence of effect heterogeneity suggested by the overall test (p < 0.0001). The distribution of the personalized effects was bimodal (Figure 1), again suggesting effect heterogeneity. However, the effects varied only in magnitude, not in direction: in the global model, FLX outperformed CBT by a factor of 0.87, with factors ranging from 0.86 to 0.89 in the personalized models.

Considering the CBT-COMB difference, COMB was more effective than CBT, as indicated by the global model (*b* = -0.25, 95% CI: -0.33 to -0.17, p < 0.0001). However, there was no indication of effect heterogeneity. This was indicated both by the minimal variance in personalized effects (-0.25 to -0.23), and by the lack of statistical evidence supporting effect heterogeneity from the overall test (p = 0.10, Table 2).

Finally, evidence of effect heterogeneity was identified for the FLX-COMB difference: COMB was more effective than FLX (*b* = -0.11, 95% CI: -0.21 to -0.02, p = 0.023), and the personalized effects ranged from -0.13 to -0.10. The overall test for effect heterogeneity produced a low p-value (p < 0.0001). The personalized effects also exhibited a bimodal distribution (Figure 1), albeit less pronounced than that found for the CBT-FLX difference. Again, the effects varied only in magnitude: in the global model, COMB outperformed FLX by a factor of 0.89, with factors ranging from 0.88 to 0.91 in the personalized models.

Importance of patient characteristics

For the between-treatment differences that exhibited effect heterogeneity, the variable importance measures of the patient characteristics are shown in Figure 2. Considering the CBT-FLX difference, baseline depression severity was the only patient characteristic with large variable importance. In contrast, three patient characteristics were prominent in the FLX-COMB difference, again including baseline depression severity, but also the patients' treatment expectations and childhood trauma.

Functional forms of effect modifications

For patient characteristics deemed important, partial dependence plots were generated to uncover the functional forms of the effect modifications. As can be seen in the top-left corner of Figure 3, FLX's superiority over CBT was more pronounced with more severe baseline depression. In contrast, COMB's superiority over FLX was reduced with more severe baseline depression (Figure 3, top-right). In addition, COMB's superiority over FLX was stronger in patients who reported higher treatment expectations, whereas childhood trauma tended to reduce the difference slightly (Figure 3, bottom-left and –right). These impacts of baseline depression severity, treatment expectations, and childhood trauma on COMB's superiority over FLX were also evident when considering the three characteristics simultaneously (Figure 4).

Discussion

Our aim was to estimate personalized treatment advantages observed in the acute phase of the TADS trial as a function of 23 patient baseline characteristics. The most over-riding finding is that combined treatment with cognitive-behavioural therapy and fluoxetine was consistently superior to either cognitive-behavioural therapy or fluoxetine administered alone, despite some heterogeneity in these effects. Thus, rather than supporting the need for specific matchings between patients and treatments, as demanded by personalized medicine, the personalized treatment advantages that we estimated suggest that the combined treatment should be a preferred treatment across a large array of patient characteristics, including socioeconomic, demographic and clinical characteristics, and childhood experiences. Previous analyses of the TADS data had already uncovered that the combined treatment was the most effective in average (March et al., 2004) and had the best benefit-harm profile (March et al., 2007). Our results additionally suggest that the combined treatment's superior effectiveness might hold for a broad range of patients.

The last claim is bolstered by our analysis' ability to incorporate 23 potential effect modifiers with arbitrary combinations. This was possible due to the model-based random forest's ability to take into account low- as well as high-dimensional effect modification. The superiority of the combined treatment should, therefore, generalize across a diverse spectrum of patients, characterized by various combinations of these baseline characteristics. Furthermore, TADS was designed to be an effectiveness trial (Hollon et al., 2005; March et al., 2004; TADS Team, 2005). Accordingly, it featured a relatively large sample that was quite heterogeneous, mirroring the heterogeneity of patients seen in actual clinical practice (Hollon et al., 2005; March et al., 2004; TADS Team, 2005). As such, its data was believed ripe for effect-modification analysis (TADS Team, 2003, 2005), and it was suggested that the TADS results should be broadly applicable to youths seeking treatment for depression across the USA (Hollon et al., 2005; March et al., 2004; TADS Team, 2005). The model-based random forest's ability to investigate patient heterogeneity comprehensively is clearly appropriate for such a sample.

A second finding was that fluoxetine was superior to cognitive-behavioural therapy, consistent with the original effectiveness analysis published by the TADS team (March et al., 2004). Evidence of effect modification was also present in this treatment advantage, however, the personalized treatment advantages varied again only in magnitude, but not in direction. The heterogeneity was driven by baseline depression severity: with severer forms of depression, the superiority of fluoxetine was more pronounced, whereas for adolescents with mild to moderate depression, the difference between the treatments diminished. This seems to be in line with findings in depressed adults that SSRI treatment is relatively more effective with higher baseline severity of depression (Fournier et al., 2010; Kirsch et al., 2008), although this finding has been disputed (Fountoulakis et al., 2013; Rabinowitz et al., 2016). Future research should further clarify this issue in depressed adolescents.

Notwithstanding the statistical comprehensiveness of our analysis, it is surprising that we did not find evidence that allowed for an unequivocal personalization of treatments. There are several potential explanations. First, the patient characteristics included in our analyses did not encompass a variety of other potential effect modifiers (Iosifescu, 2011; Leuchter et al., 2010; McGrath et al., 2013; Strawbridge et al., 2017; Uhr et al., 2008), in particular a variety of potential biomarkers (Harmer et al., 2011; Leuchter et al., 2016; Porcelli et al., 2011). Eventually, one or several of these variables would produce effect modifications that allow for personalizing treatment decisions (e.g. Uhr et al., 2008). Future studies should close this gap especially concerning the role of biological measures.

Second, commonly-used depression rating scales have been criticized for not being psychometrically valid (Bech, 2010; Fried and Nesse, 2015; Fried et al., 2016) and this issue has been raised with the CDRS-R (Isa et al., 2014). If the outcome measure used is unreliable, true effect heterogeneity might be clouded by measurement error, especially in case of smaller effect modifications. Note, however, that this is unlikely to explain all of our results, since our analysis did reveal some effect heterogeneity.

Third, RCTs, including TADS, are usually powered to detect average treatment effects. As a result, sample sizes within subgroups, as implied by effect modifiers, often become small and imbalanced (Rothwell, 2005; Willke et al., 2012). Several patient characteristics in our analysis might have been affected by this problem, such as dysthymia, ADHD, and some forms of childhood trauma (see table 1). For example, only 46 TADS patients had dysthymia, so that the number of dysthymic patients in the four groups ranged from a low of six to a high of 17. Any effect-modification analysis is limited by these low numbers, making the results less reliable for these characteristics. Note that this problem did not universally apply to the patient characteristics included in our analysis (table 1), however.

Fourth, it might be necessary to take harms additionally into account for arriving at personalized treatment decisions. As already mentioned above, although we found a consistent treatment advantage of fluoxetine over cognitive-behavioral therapy, there was heterogeneity in this advantage. In fact, the personalized advantages showed a bi-modal distribution with the bi-modality being driven by baseline depression severity: with less severe depression the advantage diminished. If fluoxetine's superiority diminishes for less severely depressed adolescents, cognitive-behavioural therapy might be a better treatment choice for these patients if the (personalized) risk of harm associated with fluoxetine is additionally taken into account, because SSRI treatment was found to confer some danger for adolescents especially regarding suicidality (Sharma et al., 2016). This resonates indeed with the conclusion drawn by Cipriani et al (2016) who - based on their network meta-analysis - stated that fluoxetine should only be considered for adolescents with moderate-severe depression. Future studies should therefore examine more closely the trade-off between personalized treatment advantage and personalized harm. Note, however, that the combined treatment of fluoxetine and cognitive-behavioral therapy was found to have the best benefit-harm account in the TADS sample (March et al., 2007), so this line of reasoning seems less important for the combined therapy, at least in the TADS sample.

Fifth, the previous literature has produced mixed results regarding the effectiveness of both combined treatments (Brent et al., 2008; Cox et al., 2014; Dubicka et al., 2010; Goodyer et al., 2008; Hetrick et al., 2011; Ma et al., 2014; Melvin et al., 2006) and SSRIs (Bridge et al., 2007; Cheung et al., 2008; Cipriani et al., 2016; Cox et al., 2012; Emslie et al., 2004; Emslie et al., 2008; Findling et al., 2013; Hetrick et al., 2012; Le Noury et al., 2015, 2016; Masi et al., 2010; Usala et al., 2008; Wagner et al., 2003; Wagner et al., 2006; Wagner et al., 2004; Whittington et al., 2004) in adolescents; though the drug that has received the most-consistent empirical support has been fluoxetine (Cipriani et al., 2016; Masi et al., 2010; Usala et al., 2008). In addition, the limited effectiveness of cognitive-behavioural therapy in TADS was unanticipated (Harrington et al., 1998; Hollon et al., 2005; Klein et al., 2007; Masi et al., 2010; Reinecke et al., 1998; Weersing and Brent, 2006; Weersing et al., 2017; Weersing et al., 2009; Weisz et al., 2006). With such mixed results, two scenarios are plausible. On one hand, this situation might be caused by the presence of strong effect modifiers: if a treatment has a positive effect in some patients, but no or a negative effect in others, the average effect is small or zero. On the other hand, the treatments might be limited in what they can achieve and effect modification may be limited accordingly.

Unfortunately, whereas our results seem more in line with the second scenario, the broader state of affairs is less clear. Existing studies are difficult to compare, since they refer to a variety of comparisons, including those comparing active treatment and placebo (Cheung et al., 2010; Curry et al., 2006; Emslie et al., 2012), comparisons between different forms of psychotherapy (Birmaher et al., 2000; Brent et al., 1998), comparisons of psychotherapy against inactive control conditions (Harrington et al., 1998), and comparisons of combined treatments with medication augmentation and alterations, in adolescents with SSRI-resistant depression (Asarnow et al., 2009). In addition, the studies were limited by the drawbacks of conventional effect-modification analyses. Not surprisingly, these studies failed to reveal patient characteristics that show up consistently as effect modifiers. On the other hand, the shortage of robust effect modifiers in previously-published literature agrees with our findings that only three out of the 23 patient characteristics we assessed acted as an effect modifier, and that the effect heterogeneity induced by these modifiers was limited.

Overall, then, we found evidence for the superiority of the combination of cognitive-behaviour therapy and fluoxetine in the TADS sample. Due to the issues discussed above, our results should nevertheless be considered exploratory. Future studies must replicate and validate our findings in even larger trials, or in data sets combining several trials, and include important additional patient characteristics.

Limitations

Limitations of the TADS trial have been noted by others (Curry et al., 2006; March et al., 2004) and other issues were addressed above. Additional misgivings should be mentioned. First, analyses were done separately for every treatment-treatment pair. This might have introduced additional variability. Technically, it is possible to analyze a global model that contains all treatment conditions simultaneously (Seibold et al., 2018a). However, the results of such an analysis would not be straightforward to interpret. Second, other statistical approaches exist that allow for a more comprehensive assessment of effect modification than the conventional approaches based on stratification and parametric interaction effects. These include, for example, recently-proposed machine-learning approaches (Chekroud et al., 2016; Jakubovski and Bloch, 2014; Kelly et al., 2015) and the multivariable risk prediction approach (Kent et al., 2010; Perlis, 2013; Varadhan et al., 2013). A systematic comparison of these methods – including the model-based random forest – has not been carried out yet, especially concerning the pivotal question of whether these approaches would lead to divergent treatment decisions. Future studies should address this issue. Finally, our analysis was a *post-hoc* analysis of data collected during a trial that was not originally designed for our intended analysis.

Conclusions

In the TADS sample, the combination of cognitive-behaviour therapy and fluoxetine was superior to either therapy used alone across a broad range of patient characteristics. Future studies may benefit from applying the model-based random forest method to analyse effect heterogeneity.

Conflicts of interest

None

Role of funding source

Heidi Seibold and Torsten Hothorn were financially supported by the Swiss National Science Foundation for carrying out work related to this manuscript [205321_163456]. The funding agency did not have any role in study design, in collection, analysis or interpretation of data, in writing the manuscript, or in the decision to submit the manuscript for publication.

Acknowledgements

Data used in the preparation of this manuscript were obtained and analysed from the controlled access datasets distributed from the National Institute of Mental Health (NIMH)-supported National Database for Clinical Trials (NDCT, https://data-archive.nimh.nih.gov/ndct/), located in the United States of America (USA). The NDCT is a collaborative informatics system created by the NIMH to provide a national resource to support and accelerate discovery related to clinical trial research in mental health. Dataset identifier: Clinical Trials #2145. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIMH or of those submitting original data to the NDCT.

References

Alemayehu, D., Chen, Y., Markatou, M., 2018. A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. Statistical methods in medical research 27(12), 3658-3678.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV). American Psychiatric Association, Washington, DC.

Asarnow, J.R., Emslie, G., Clarke, G., Wagner, K.D., Spirito, A., Vitiello, B., Iyengar, S., Shamseddeen, W., Ritz, L., McCracken, J., Strober, M., Suddath, R., Leonard, H., Porta, G., Keller, M., Brent, D., 2009. Treatment of selective serotonin reuptake inhibitor-resistant depression in adolescents: predictors and moderators of treatment response. Journal of the American Academy of Child and Adolescent Psychiatry 48(3), 330-339.

Bech, P., 2010. Is the antidepressive effect of second-generation antidepressants a myth? Psychological medicine 40(2), 181-186.

Birmaher, B., Brent, D.A., Kolko, D., Baugher, M., Bridge, J., Holder, D., Iyengar, S., Ulloa, R.E., 2000. Clinical outcome after short-term psychotherapy for adolescents with major depressive disorder. Archives of general psychiatry 57(1), 29-36.

Brent, D., Emslie, G., Clarke, G., Wagner, K.D., Asarnow, J.R., Keller, M., Vitiello, B., Ritz, L., Iyengar, S., Abebe, K., Birmaher, B., Ryan, N., Kennard, B., Hughes, C., DeBar, L., McCracken, J., Strober, M., Suddath, R., Spirito, A., Leonard, H., Melhem, N., Porta, G., Onorato, M., Zelazny, J., 2008. Switching to another SSRI or to venlafaxine with or without cognitive behavioral therapy for adolescents with SSRI-resistant depression: the TORDIA randomized controlled trial. Jama 299(8), 901-913.

Brent, D.A., Kolko, D.J., Birmaher, B., Baugher, M., Bridge, J., Roth, C., Holder, D., 1998. Predictors of treatment efficacy in a clinical trial of three psychosocial treatments for adolescent depression. Journal of the American Academy of Child and Adolescent Psychiatry 37(9), 906-914.

Bridge, J.A., Iyengar, S., Salary, C.B., Barbe, R.P., Birmaher, B., Pincus, H.A., Ren, L., Brent, D.A., 2007. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: a meta-analysis of randomized controlled trials. Jama 297(15), 1683-1696.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry 3(3), 243-250.

Cheung, A., Kusumakar, V., Kutcher, S., Dubo, E., Garland, J., Weiss, M., Kiss, A., Levitt, A., 2008. Maintenance study for adolescent depression. Journal of child and adolescent psychopharmacology 18(4), 389-394.

Cheung, A., Mayes, T., Levitt, A., Schaffer, A., Michalak, E., Kiss, A., Emslie, G., 2010. Anxiety as a predictor of treatment outcome in children and adolescents with depression. Journal of child and adolescent psychopharmacology 20(3), 211-216.

Cipriani, A., Zhou, X., Del Giovane, C., Hetrick, S.E., Qin, B., Whittington, C., Coghill, D., Zhang, Y., Hazell, P., Leucht, S., Cuijpers, P., Pu, J., Cohen, D., Ravindran, A.V., Liu, Y., Michael, K.D., Yang, L., Liu, L., Xie, P., 2016. Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. Lancet (London, England) 388(10047), 881-890.

Cox, G.R., Callahan, P., Churchill, R., Hunot, V., Merry, S.N., Parker, A.G., Hetrick, S.E., 2014. Psychological therapies versus antidepressant medication, alone and in combination for depression in children and adolescents. The Cochrane database of systematic reviews 11, CD008324-CD008324.

Cox, G.R., Fisher, C.A., De Silva, S., Phelan, M., Akinwale, O.P., Simmons, M.B., Hetrick, S.E., 2012. Interventions for preventing relapse and recurrence of a depressive disorder in children and adolescents. The Cochrane database of systematic reviews 11, CD007504.

Curry, J., Rohde, P., Simons, A., Silva, S., Vitiello, B., Kratochvil, C., Reinecke, M., Feeny, N., Wells, K., Pathak, S., Weller, E., Rosenberg, D., Kennard, B., Robins, M., Ginsburg, G., March, J., 2006. Predictors and moderators of acute outcome in the Treatment for Adolescents with Depression Study (TADS). Journal of the American Academy of Child and Adolescent Psychiatry 45(12), 1427-1439.

Dahabreh, I.J., Hayward, R., Kent, D.M., 2016. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International journal of epidemiology 45(6), 2184-2193.

Dubicka, B., Elvins, R., Roberts, C., Chick, G., Wilkinson, P., Goodyer, I.M., 2010. Combined treatment with cognitive-behavioural therapy in adolescent depression: meta-analysis. The British journal of psychiatry : the journal of mental science 197(6), 433-440.

Efron, B., Hastie, T., 2016. Computer age statistical inference. Algorithms, evidence, and data science. Cambridge University Press, New York.

Emslie, G.J., Heiligenstein, J.H., Hoog, S.L., Wagner, K.D., Findling, R.L., McCracken, J.T., Nilsson, M.E., Jacobson, J.G., 2004. Fluoxetine treatment for prevention of relapse of depression in children and adolescents: a double-blind, placebo-controlled study. Journal of the American Academy of Child and Adolescent Psychiatry 43(11), 1397-1405.

Emslie, G.J., Kennard, B.D., Mayes, T.L., Nakonezny, P.A., Zhu, L., Tao, R., Hughes, C., Croarkin, P., 2012. Insomnia moderates outcome of serotonin-selective reuptake inhibitor treatment in depressed youth. Journal of child and adolescent psychopharmacology 22(1), 21-28.

Emslie, G.J., Kennard, B.D., Mayes, T.L., Nightingale-Teresi, J., Carmody, T., Hughes, C.W., Rush, A.J., Tao, R., Rintelmann, J.W., 2008. Fluoxetine versus placebo in preventing relapse of major depression in children and adolescents. The American journal of psychiatry 165(4), 459-467.

Findling, R.L., Robb, A., Bose, A., 2013. Escitalopram in the treatment of adolescent depression: a randomized, double-blind, placebo-controlled extension trial. Journal of child and adolescent psychopharmacology 23(7), 468-480.

Fountoulakis, K.N., Veroniki, A.A., Siamouli, M., Moller, H.J., 2013. No role for initial severity on the efficacy of antidepressants: results of a multi-meta-analysis. Annals of general psychiatry 12(1), 26.

Fournier, J.C., DeRubeis, R.J., Hollon, S.D., Dimidjian, S., Amsterdam, J.D., Shelton, R.C., Fawcett, J., 2010. Antidepressant drug effects and depression severity: a patient-level meta-analysis. Jama 303(1), 47-53.

Fried, E.I., Nesse, R.M., 2015. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC medicine 13(1), 72.

Fried, E.I., van Borkulo, C.D., Epskamp, S., Schoevers, R.A., Tuerlinckx, F., Borsboom, D., 2016. Measuring Depression Over Time . . . or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression. Psychological assessment 28(11), 1354-1367.

Goodyer, I.M., Dubicka, B., Wilkinson, P., Kelvin, R., Roberts, C., Byford, S., Breen, S., Ford, C., Barrett, B., Leech, A., Rothwell, J., White, L., Harrington, R., 2008. A randomised controlled trial of cognitive behaviour therapy in adolescents with major depression treated by selective serotonin reuptake inhibitors. The ADAPT trial. Health technology assessment (Winchester, England) 12(14), iii-iv, ix-60.

Harmer, C.J., Cowen, P.J., Goodwin, G.M., 2011. Efficacy markers in depression. Journal of psychopharmacology (Oxford, England) 25(9), 1148-1158.

Harrington, R., Whittaker, J., Shoebridge, P., Campbell, F., 1998. Systematic review of efficacy of cognitive behaviour therapies in childhood and adolescent depressive disorder. BMJ (Clinical research ed.) 316(7144), 1559-1563.

Hetrick, S.E., Cox, G.R., Merry, S.N., 2011. Treatment-resistant depression in adolescents: is the addition of cognitive behavioral therapy of benefit? Psychology research and behavior management 4, 97-112.

Hetrick, S.E., McKenzie, J.E., Cox, G.R., Simmons, M.B., Merry, S.N., 2012. Newer generation antidepressants for depressive disorders in children and adolescents. The Cochrane database of systematic reviews 11, CD004851.

Hollon, S.D., Garber, J., Shelton, R.C., 2005. Treatment of depression in adolescents with cognitive behavior therapy and medications: A commentary on the TADS project. Cognitive and Behavioral Practice 12(2), 149-155.

Iosifescu, D.V., 2011. Electroencephalography-derived biomarkers of antidepressant response. Harv Rev Psychiatry 19(3), 144-154.

Isa, A., Bernstein, I., Trivedi, M.H., Mayes, T.L., Kennard, B., Emslie, G., 2014. Childhood Depression Subscales Using Repeated Sessions on Children's Depression Rating Scale – Revised (CDRS-R) Scores. Journal of child and adolescent psychopharmacology 24(6), 318-324.

Jakubovski, E., Bloch, M.H., 2014. Prognostic subgroups for citalopram response in the STAR*D trial. The Journal of clinical psychiatry 75(7), 738-747.

Kelly, J.M., Jakubovski, E., Bloch, M.H., 2015. Prognostic subgroups for remission and response in the Coordinated Anxiety Learning and Management (CALM) trial. The Journal of clinical psychiatry 76(3), 267-278.

Kent, D.M., Rothwell, P.M., Ioannidis, J.P., Altman, D.G., Hayward, R.A., 2010. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 11, 85.

Kessler, R.C., van Loo, H.M., Wardenaar, K.J., Bossarte, R.M., Brenner, L.A., Ebert, D.D., de Jonge, P., Nierenberg, A.A., Rosellini, A.J., Sampson, N.A., Schoevers, R.A., Wilcox, M.A., Zaslavsky, A.M., 2017. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiology and psychiatric sciences 26(1), 22-36.

Kirsch, I., Deacon, B.J., Huedo-Medina, T.B., Scoboria, A., Moore, T.J., Johnson, B.T., 2008. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS medicine 5(2), e45-e45.

Klein, J.B., Jacobs, R.H., Reinecke, M.A., 2007. Cognitive-behavioral therapy for adolescent depression: a meta-analytic investigation of changes in effect-size estimates. Journal of the American Academy of Child and Adolescent Psychiatry 46(11), 1403-1413.

Kratochvil, C.J., May, D.E., Silva, S.G., Madaan, V., Puumala, S.E., Curry, J.F., Walkup, J., Kepley, H., Vitiello, B., March, J.S., 2009. Treatment response in depressed adolescents with and without co-morbid attention-deficit/hyperactivity disorder in the Treatment for Adolescents with Depression Study. Journal of child and adolescent psychopharmacology 19(5), 519-527.

Le Noury, J., Nardo, J.M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., Abi-Jaoude, E., 2015. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. BMJ (Clinical research ed.) 351, h4320.

Le Noury, J., Nardo, J.M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., Abi-Jaoude, E., 2016. Study 329 continuation phase: Safety and efficacy of paroxetine and imipramine in extended treatment of adolescent major depression. The International journal of risk & safety in medicine 28(3), 143-161.

Leuchter, A.F., Cook, I.A., Hamilton, S.P., Narr, K.L., Toga, A., Hunter, A.M., Faull, K., Whitelegge, J., Andrews, A.M., Loo, J., Way, B., Nelson, S.F., Horvath, S., Lebowitz, B.D., 2010. Biomarkers to predict antidepressant response. Current psychiatry reports 12(6), 553-562.

Leuchter, A.F., Hunter, A.M., Jain, F.A., Tartter, M., Crump, C., Cook, I.A., 2016. Escitalopram but not placebo modulates brain rhythmic oscillatory activity in the first week of treatment of Major Depressive Disorder. Journal of psychiatric research 84, 174-183.

Lewis, C.C., Simons, A.D., Nguyen, L.J., Murakami, J.L., Reid, M.W., Silva, S.G., March, J.S., 2010. Impact of childhood trauma on treatment outcome in the Treatment for Adolescents with Depression Study (TADS). Journal of the American Academy of Child and Adolescent Psychiatry 49(2), 132-140.

Ma, D., Zhang, Z., Zhang, X., Li, L., 2014. Comparative efficacy, acceptability, and safety of medicinal, cognitive-behavioral therapy, and placebo treatments for acute major depressive disorder in children and adolescents: a multiple-treatments meta-analysis. Current medical research and opinion 30(6), 971-995.

March, J., Silva, S., Petrycki, S., Curry, J., Wells, K., Fairbank, J., Burns, B., Domino, M., McNulty, S., Vitiello, B., Severe, J., 2004. Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents With Depression Study (TADS) randomized controlled trial. Jama 292(7), 807-820.

March, J.S., Silva, S., Petrycki, S., Curry, J., Wells, K., Fairbank, J., Burns, B., Domino, M., McNulty, S., Vitiello, B., Severe, J., 2007. The Treatment for Adolescents With Depression Study (TADS): long-term effectiveness and safety outcomes. Archives of general psychiatry 64(10), 1132-1143.

Masi, G., Liboni, F., Brovedani, P., 2010. Pharmacotherapy of major depressive disorder in adolescents. Expert opinion on pharmacotherapy 11(3), 375-386.

Mayes, T.L., Bernstein, I.H., Haley, C.L., Kennard, B.D., Emslie, G.J., 2010. Psychometric properties of the Children's Depression Rating Scale-Revised in adolescents. Journal of child and adolescent psychopharmacology 20(6), 513-516.

McGrath, C.L., Kelley, M.E., Holtzheimer, P.E., Dunlop, B.W., Craighead, W.E., Franco, A.R., Craddock, R.C., Mayberg, H.S., 2013. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA psychiatry 70(8), 821-829.

Melvin, G.A., Tonge, B.J., King, N.J., Heyne, D., Gordon, M.S., Klimkeit, E., 2006. A comparison of cognitive-behavioral therapy, sertraline, and their combination for adolescent depression. Journal of the American Academy of Child and Adolescent Psychiatry 45(10), 1151-1161.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gotzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., Altman, D.G., 2012. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. International journal of surgery (London, England) 10(1), 28-55.

Perlis, R.H., 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biological psychiatry 74(1), 7-14.

Porcelli, S., Drago, A., Fabbri, C., Gibiino, S., Calati, R., Serretti, A., 2011. Pharmacogenetics of antidepressant response. Journal of psychiatry & neuroscience : JPN 36(2), 87-113.

Poznanski, E., Mokros, H., 1995. Children's Depression Rating Scale-Revised (CDRS-R). WPS, Los Angeles, Calif.

R Core Team, 2016. R: A Language and Environment for Statistical Computing, 3.3.2 ed. R Foundation for Statistical Computing, Vienna.

Rabinowitz, J., Werbeloff, N., Mandel, F.S., Menard, F., Marangell, L., Kapur, S., 2016. Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. The British journal of psychiatry : the journal of mental science 209(5), 427-428.

Reinecke, M.A., Ryan, N.E., DuBois, D.L., 1998. Cognitive-behavioral therapy of depression and depressive symptoms during adolescence: a review and meta-analysis. Journal of the American Academy of Child and Adolescent Psychiatry 37(1), 26-34.

Rothwell, P.M., 2005. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet (London, England) 365(9454), 176-186.

Seibold, H., Zeileis, A., Hothorn, T., 2016. Model-based recursive partitioning for subgroup analyses. International Journal of Biostatistics 12(1), 45–63.

Seibold, H., Zeileis, A., Hothorn, T., 2018a. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. Statistical Methods in Medical Research 27(10), 3104–3125.

Seibold, H., Zeileis, A., Hothorn, T., 2018b. model4you: Stratified and personalised models based on model-based trees and forests. R package version 0.9-2.

Shaffer, J.P., 1995. Multiple hypothesis testing. Annual review of psychology.

Sharma, T., Guski, L.S., Freund, N., Gotzsche, P.C., 2016. Suicidality and aggression during antidepressant treatment: systematic review and meta-analyses based on clinical study reports. BMJ (Clinical research ed.) 352, i65-i65.

Simon, G.E., Perlis, R.H., 2010. Personalized medicine for depression: can we match patients with treatments? The American journal of psychiatry 167(12), 1445-1455.

Strawbridge, R., Young, A.H., Cleare, A.J., 2017. Biomarkers for depression: recent insights, current challenges and future prospects. Neuropsychiatric disease and treatment 13, 1245-1262.

TADS Team, 2003. Treatment for Adolescents With Depression Study (TADS): rationale, design, and methods. Journal of the American Academy of Child and Adolescent Psychiatry 42(5), 531-542.

TADS Team, 2005. The Treatment for Adolescents With Depression Study (TADS): demographic and clinical characteristics. Journal of the American Academy of Child and Adolescent Psychiatry 44(1), 28-40.

Thapar, A., Collishaw, S., Pine, D.S., Thapar, A.K., 2012. Depression in adolescence. Lancet (London, England) 379(9820), 1056-1067.

Uhr, M., Tontsch, A., Namendorf, C., Ripke, S., Lucae, S., Ising, M., Dose, T., Ebinger, M., Rosenhagen, M., Kohli, M., Kloiber, S., Salyakina, D., Bettecken, T., Specht, M., Putz, B., Binder, E.B., Muller-Myhsok, B., Holsboer, F., 2008. Polymorphisms in the drug transporter gene ABCB1 predict antidepressant treatment response in depression. Neuron 57(2), 203-209.

Usala, T., Clavenna, A., Zuddas, A., Bonati, M., 2008. Randomised controlled trials of selective serotonin reuptake inhibitors in treating depression in children and adolescents: a systematic review and meta-analysis. European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology 18(1), 62-73.

Varadhan, R., Segal, J.B., Boyd, C.M., Wu, A.W., Weiss, C.O., 2013. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. Journal of clinical epidemiology 66(8), 818-825.

Wagner, K.D., Ambrosini, P., Rynn, M., Wohlberg, C., Yang, R., Greenbaum, M.S., Childress, A., Donnelly, C., Deas, D., 2003. Efficacy of sertraline in the treatment of children and adolescents with major depressive disorder: two randomized controlled trials. Jama 290(8), 1033-1041.

Wagner, K.D., Jonas, J., Findling, R.L., Ventura, D., Saikali, K., 2006. A double-blind, randomized, placebo-controlled trial of escitalopram in the treatment of pediatric depression. Journal of the American Academy of Child and Adolescent Psychiatry 45(3), 280-288.

Wagner, K.D., Robb, A.S., Findling, R.L., Jin, J., Gutierrez, M.M., Heydorn, W.E., 2004. A randomized, placebo-controlled trial of citalopram for the treatment of major depression in children and adolescents. The American journal of psychiatry 161(6), 1079-1083.

Weersing, V.R., Brent, D.A., 2006. Cognitive behavioral therapy for depression in youth. Child and adolescent psychiatric clinics of North America 15(4), 939-957, ix.

Weersing, V.R., Jeffreys, M., Do, M.T., Schwartz, K.T., Bolano, C., 2017. Evidence base update of psychosocial treatments for child and adolescent depression. Journal of clinical child and adolescent psychology : the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53 46(1), 11-43.

Weersing, V.R., Rozenman, M., Gonzalez, A., 2009. Core components of therapy in youth: do we know what to disseminate? Behavior modification 33(1), 24-47.

Weisz, J.R., McCarty, C.A., Valeri, S.M., 2006. Effects of psychotherapy for depression in children and adolescents: a meta-analysis. Psychological bulletin 132(1), 132-149.

Whittington, C.J., Kendall, T., Fonagy, P., Cottrell, D., Cotgrove, A., Boddington, E., 2004. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. Lancet (London, England) 363(9418), 1341-1345.

Willke, R.J., Zheng, Z., Subedi, P., Althin, R., Mullins, C.D., 2012. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. BMC medical research methodology 12, 185.

Zeileis, A., Hothorn, T., Hornik, K., 2008. Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17(2), 492-514.

Table 1 Overview of patient characteristics examined as effect modifiers.

Categorical patient characteristic	**Categories**	**n**	**%**	**-**	**-**	**Scale**
Gender	female	239	54.4			binary
	male	200	45.6			
Race	majority (white)	324	73.8			binary
	minority	115	26.2			
Family income	≥ 75000	100	25.4			binary
	< 75000	294	74.6			
Referral source	no advertisement	193	44			binary
	advertisement	246	56			
Dysthymia	no	392	89.5			binary
	yes	46	10.5			
Anxiety disorder	no	318	72.6			binary
	yes	120	27.4			
ADHD	no	377	85.9			binary
	yes	62	14.1			
Childhood trauma	NT	213	48.5			categorical
	TNA	148	33.7			
	CSA	38	8.7			
	PA	40	9.1			
Study site	1	20	4.6			categorical
	2	33	7.5			
	3	33	7.5			
	4	20	4.6			
	5	64	14.6			
	6	41	9.3			
	7	50	11.4			
	8	19	4.3			
	9	21	4.8			

	10	8	1.8
	11	34	7.7
	12	10	2.3
	13	86	19.6

Numeric patient characteristic	**Mean**	**SD**	**Median**	**IQR**	**Range**	
Age	14.6	1.5	15	13-16	12-17	years
Verbal intelligence	10.7	2.4	11	9-12	6-19	vocabulary subtest of the WISC-III scaled score
Current episode duration	72.2	83.1	40	20-102	3-572	weeks
Baseline depression severity	4.8	0.8	5	4-5	3-7	CGI-S score
Functional impairment	49.6	7.5	50	45-55	32-80	CGAS score
Suicidal ideation	23.7	21.8	16	7-37	0-89	SIQ-Jr score
Melancholic features	1.3	0.9	1	1-2	0-5	count
Number comorbid diagnoses	0.9	1.1	1	0-1	0-5	count
Caregiver depression	12.5	9.7	10	5-18.2	0-43	BDI-II score
Conflict with caregiver	8.4	4.8	8	5-12	0-21	CBQ score
Hopelessness	9.9	5.6	10	5.5-14.5	0-20	BHS score
Cognitive distortions	62.6	20.6	63	46-76	24-120	CNCEQ score
Treatment expectations parents	1.8	0.9	2	1-2	-3-3	7-point Likert scale
Treatment expectations adolescents	1.6	1	2	1-2	-3-3	7-point Likert scale

SD: Standard deviation; IQR: Inter-quartile range; ADHD: Attention-Deficit/Hyperactivity Disorder; NT: no trauma; TNA: trauma but no abuse; CSA: childhood sexual abuse or both sexual and physical abuse; PA: physical abuse and/or victim of a violent crime; WISC-III: Wechsler intelligence scale for children 3. Edition; CGI-S: Clinical Global Impressions-Severity scale; CGAS: Children's Global Assessment Scale; SIQ-Jr: Suicidal Ideation Questionnaire Grades 7–9; BDI-II: Beck Depression Inventory II; CBQ: Conflict Behaviour Questionnaire; BHS: Beck Hopelessness Scale; CNCEQ: Children`s Negative Cognitive Errors Questionnaire.

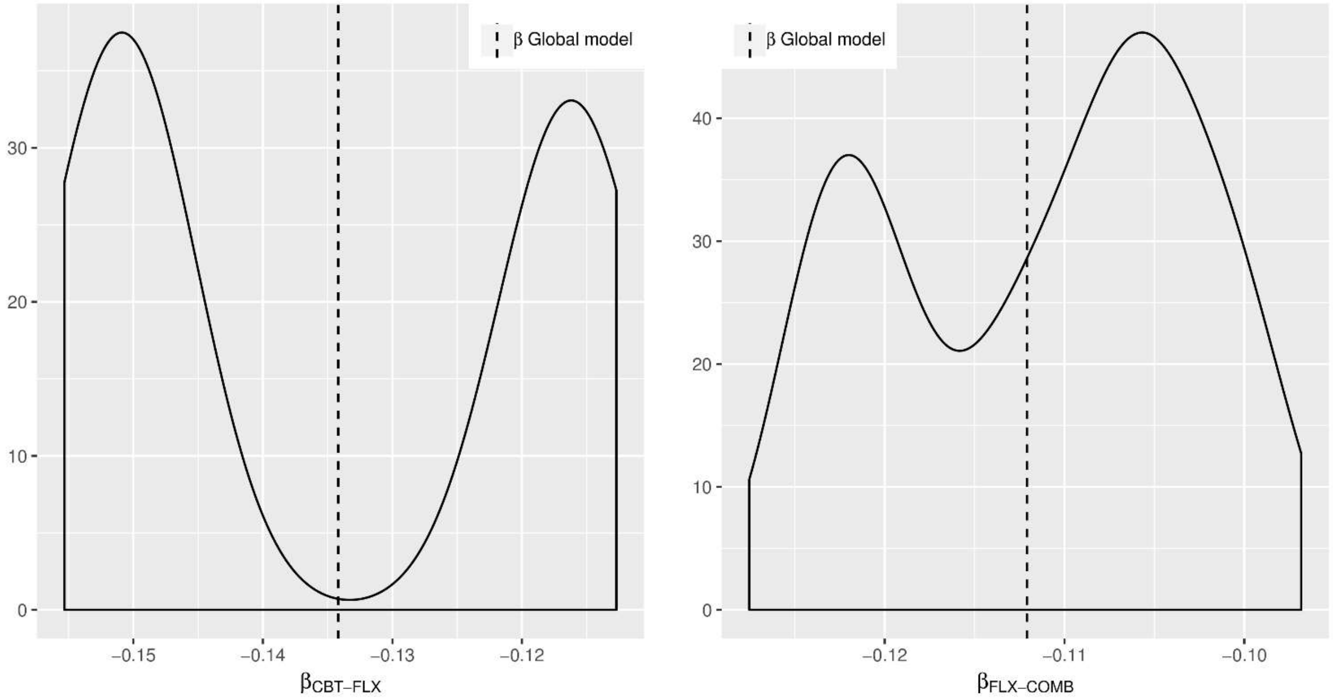
Table 2. Global model and variability of patient-specific effects.

		Global model				**Variability of model coefficients across patient characteristics**					**Test for overall effect heterogeneity** [a]
Treatment-treatment difference		*Coefficient*	*95% CI low*	*95% CI high*	*p-value*	*Median*	*IQR low*	*IQR high*	*Min*	*Max*	*p-value*
1	Intercept	-0.37	-0.42	-0.31	< 0.0001	-0.36	-0.37	-0.36	-0.37	-0.36	
	CBT vs. FLX	-0.13	-0.22	-0.05	0.0040	-0.15	-0.15	-0.12	-0.16	-0.11	< 0.0001
2	Intercept	-0.37	-0.42	-0.32	< 0.0001	-0.37	-0.37	-0.37	-0.37	-0.36	
	CBT vs. COMB	-0.25	-0.33	-0.17	< 0.0001	-0.24	-0.25	-0.24	-0.25	-0.23	0.10
3	Intercept	-0.50	-0.57	-0.44	< 0.0001	-0.51	-0.51	-0.48	-0.52	-0.48	
	FLX vs. COMB	-0.11	-0.21	-0.02	0.023	-0.11	-0.12	-0.10	-0.13	-0.10	< 0.0001

CBT: Cognitive-behavioural therapy. FLX: Fluoxetine. COMB: Combination of CBT and FLX. CI: Confidence interval. IQR: Interquartile range; Min: Minimum; Max: Maximum.

[a] Tests whether patient-specific treatment-treatment-differences result in an improvement of the global model.

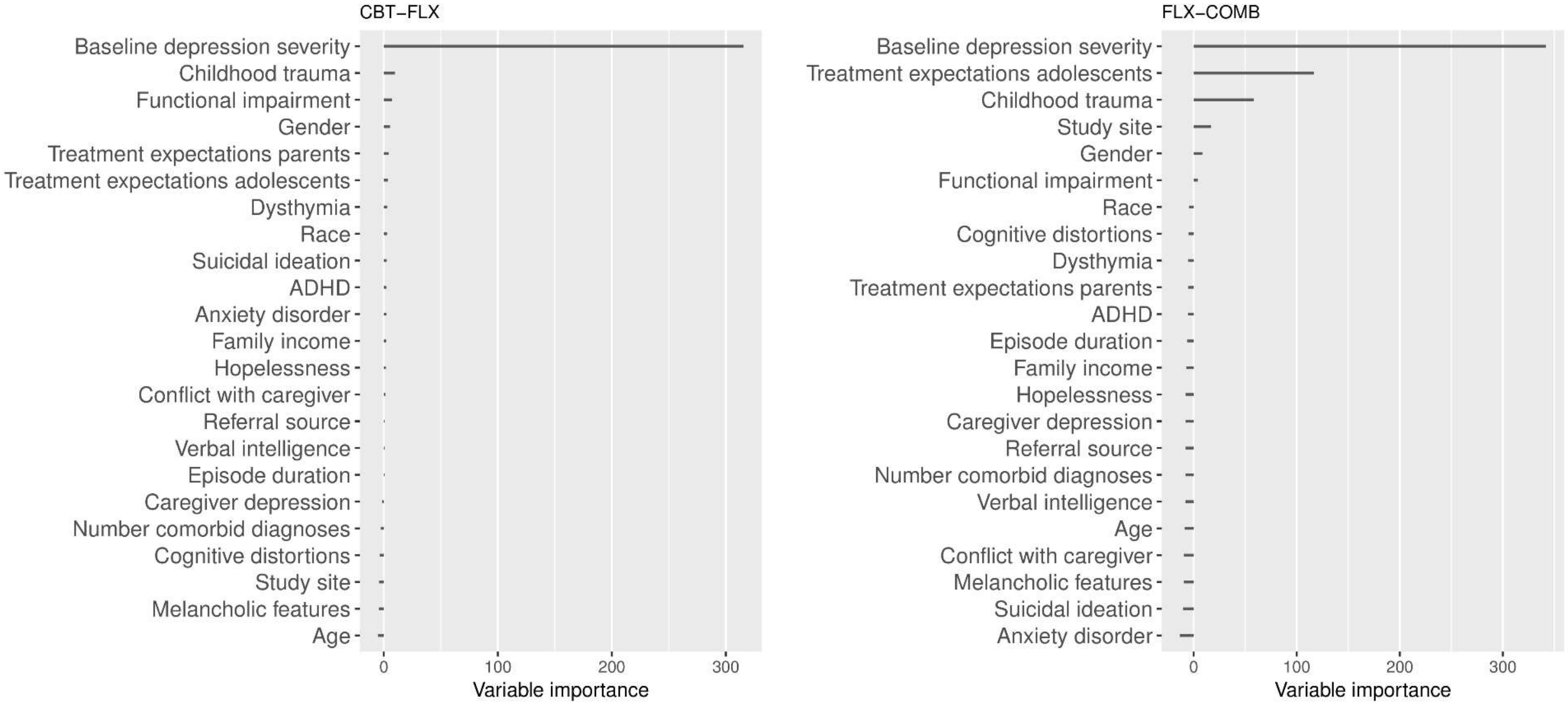
Figure 1. Patient-specific inter-treatment-differences



CBT: Cognitive-behavioural therapy; FLX: Fluoxetine; COMB: Combination of CBT and FLX; β: regression coefficient (inter-treatment-difference)

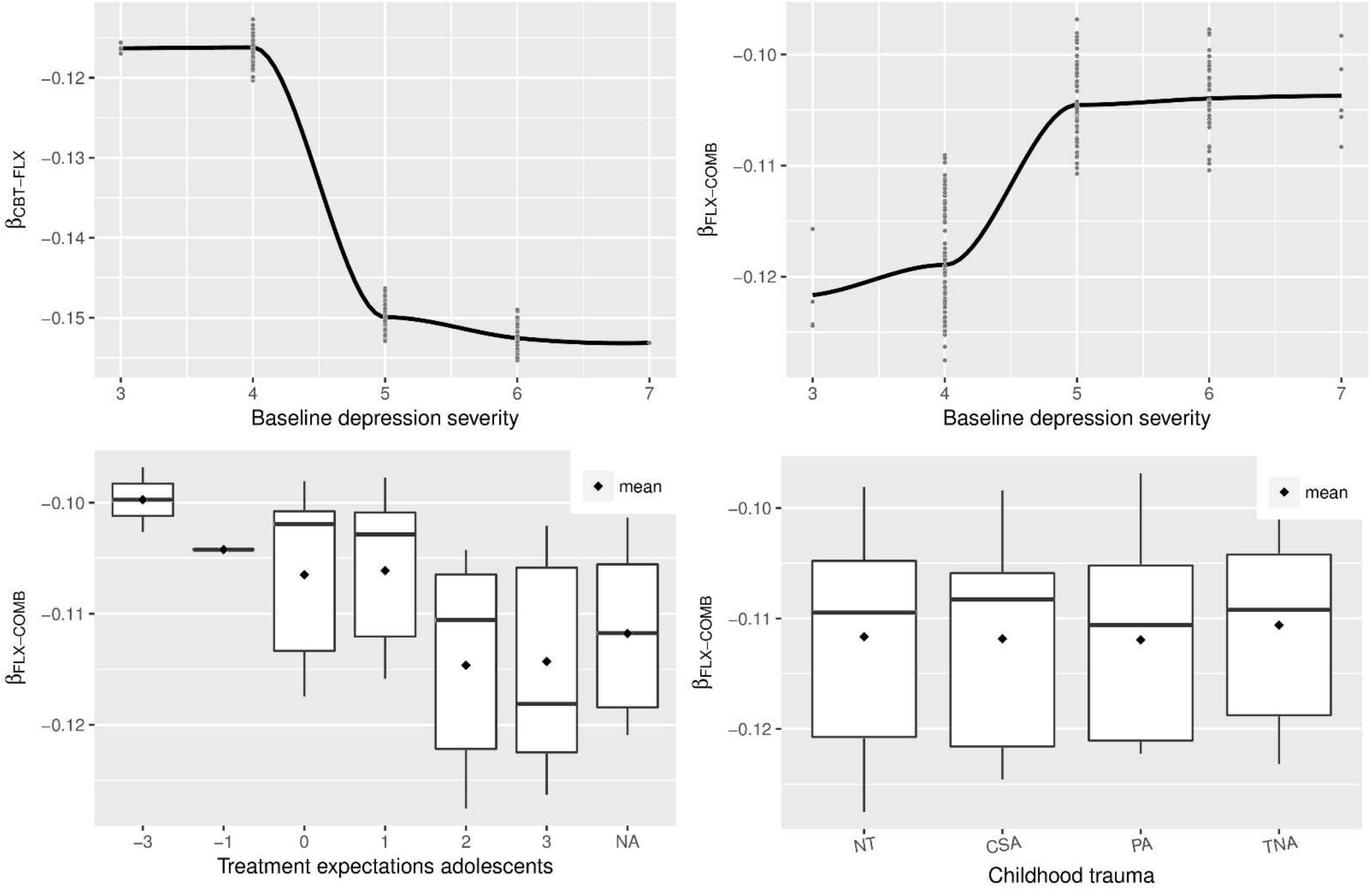
The figure shows kernel density plots of patient-specific inter-treatment-differences as estimated by model-based random forests, along with the global inter-treatment-difference (“Global model”). The left plot shows patient-specific differences between CBT and FLX, whereas the right plot shows the patient-specific differences between FLX and COMB.

Figure 2. Variable importance of patient characteristics examined as effect modifiers.



CBT: Cognitive-behavioural therapy; FLX: Fluoxetine; COMB: Combination of CBT and FLX; ADHD: Attention-deficit/hyperactivity disorder. The figure shows the variable importance of patient characteristics that might act as effect modifiers of inter-treatment-differences, as estimated by model-based random forests. Patient characteristics with a high variable importance are likely to be effect modifiers. The left plot shows patient characteristics that modify the inter-treatment-difference of CBT and FLX, whereas the right plot is for the inter-treatment-difference of FLX and COMB.

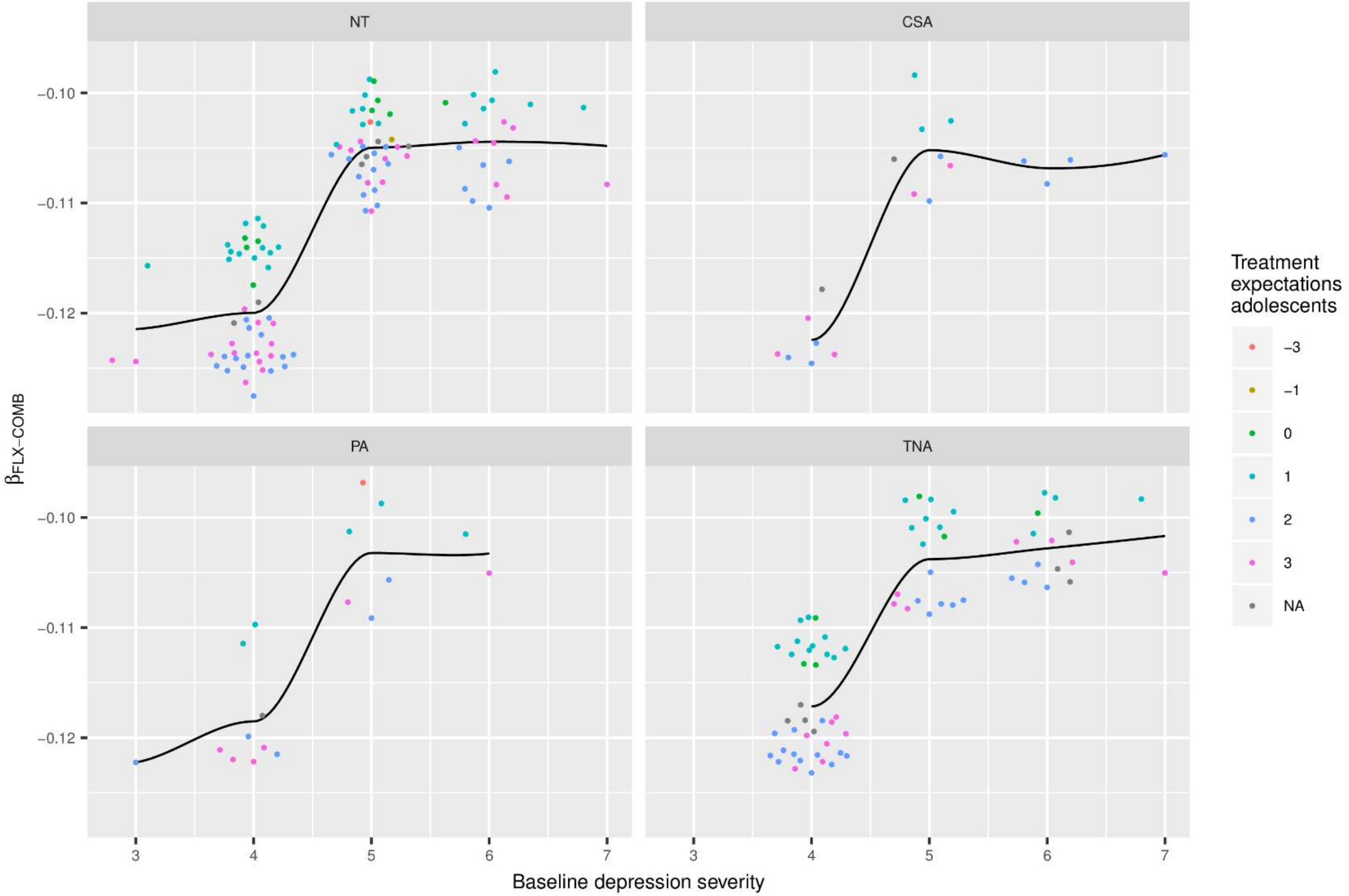
Figure 3. Patient-specific inter-treatment differences as a function of effect modifiers



CBT: Cognitive-behavioural therapy; FLX: Fluoxetine; COMB: Combination of CBT and FLX; β: regression coefficient (inter-treatment-difference); NA: Not available (missing values); NT: no trauma; CSA: childhood sexual abuse or both sexual and physical abuse; PA: physical abuse and/or victim of a violent crime; TNA: trauma but no abuse.

The figure shows partial dependence plots displaying patient-specific inter-treatment differences as a function of effect modifiers. The plot in the top left corner shows the inter-treatment difference of CBT and FLX as a function of baseline depression severity, whereas the remaining plots show the inter-treatment difference of FLX and COMB as a function of baseline depression severity, adolescents' treatment expectations, and childhood trauma, respectively. Baseline depression severity was measured via the Clinical Global Impressions-Severity scale (CGI-S): 3 = “mildly ill“; 4 = “moderately ill“; 5 = “markedly ill”; 6 = “severely ill”; 7 = “among the most extremely ill patients”.

Figure 4. Simultaneous impact of baseline depression severity, treatment expectations, and childhood trauma on the superiority of the combination of fluoxetine with cognitive-behavioural therapy over fluoxetine alone



FLX: Fluoxetine; COMB: Combination of cognitive-behavioural therapy and FLX; β: regression coefficient (inter-treatment-difference); NA: Not available (missing values); NT: no trauma; CSA: childhood sexual abuse or both sexual and physical abuse; PA: physical abuse and/or victim of a violent crime; TNA: trauma but no abuse.

The figure shows the impact of baseline depression severity, treatment expectations, and childhood trauma on COMB's superiority over FLX. The figure is divided into four subfigures, each one portraying one trauma category (NT, CSA, PA, TNA). In each subfigure, the personalized treatment-differences are plotted against baseline depression severity and color-coded for treatment expectations. Note that the subfigures were jittered to reduce over-plotting: within each level of baseline depression severity, the corresponding points would truly lie on a single line (see Figure 3, top-right plot) rather than being scattered around the severity level. Baseline depression severity was measured via the Clinical Global Impressions-Severity scale (CGI-S): 3 = "mildly ill“; 4 = "moderately ill“; 5 = "markedly ill"; 6 = "severely ill"; 7 = "among the most extremely ill patients".