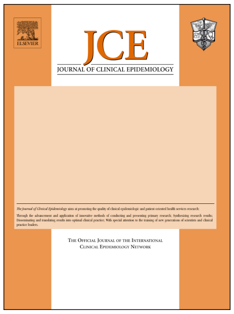


Large responses to antidepressants or methodological artifacts? A secondary analysis of STAR*D, a single-arm, open-label, non-industry antidepressant trial

Colin Xu, Florian Naudet, Thomas T. Kim, Michael P. Hengartner, Mark A. Horowitz, Irving Kirsch, Joanna Moncrieff, Ed Pigott, Martin Plöderl

PII: S0895-4356(25)00276-8

DOI: https://doi.org/10.1016/j.jclinepi.2025.111943

Reference: JCE 111943

To appear in: *Journal of Clinical Epidemiology*

Received Date: 19 September 2024

Revised Date: 17 June 2025

Accepted Date: 19 August 2025

Large responses to antidepressants or methodological artifacts?

A secondary analysis of STAR*D, a single-arm, open-label, non-industry antidepressant trial

Colin Xu,[1] Florian Naudet,[2,3] Thomas T. Kim,[4] Michael P. Hengartner,[5] Mark A. Horowitz,[6]
Irving Kirsch,[8] Joanna Moncrieff,[6,7] Ed Pigott,[9] Martin Plöderl [10] *

[1] Department of Psychology and Communication, Institute for Modeling Collaboration and Innovation, University of Idaho, Moscow, Idaho, USA

[2] Univ Rennes, CHU Rennes, Inserm, EHESP, Irset (Institut de Recherche en Santé Environnement Et Travail)-UMR_S 1085, CIC 1414 [(Centre d'investigation clinique de Rennes)], F- 35000, Rennes, France.

[3] Institut Universitaire de France (IUF), Paris, France.

[4] Department of Psychiatry, Weill Cornell Medical College, New York, NY

[5] Department of Applied Psychology, Kalaidos University of Applied Sciences, Zurich, Switzerland.

[6] Research and Development Department, Research and Development Department, North East London NHS Foundation Trust (NELFT), Rainham, UK.

[7] Division of Psychiatry, University College London, London, UK.

[8] Harvard Medical School, Arlington, Massachusetts, USA.

[9] None, Wakefield, Rhode Island, USA.

[10] Department of Inpatient Psychotherapy and Crisis Intervention, University Clinic for Psychiatry, Psychotherapy, and Psychosomatics, Christian Doppler Clinic, Paracelsus Medical University, Salzburg, Austria.

*Corresponding Author: Correspondence Adress: Martin Plöderl, Christian Doppler Klinik, Bereich für Stationäre Psychotherapie und Krisenintervention, Ignaz-Harrer-Str. 79, 5020 Salzburg, Austria. Email: m.ploederl@salk.at, Telephone: +43 5 7255 56679

Abstract

Objectives: To replicate Stone et al.'s (2022) [1] finding that the distribution of response in clinical antidepressant trials is trimodal with large, medium-effect, and small subgroups.

Methods: To apply finite mixture modeling to pre-post Hamilton Depression Rating Scale (HDRS) differences (n = 2184) of STAR*D study's level 1, a single-arm, open-label study. For a successful replication, the best fitting model had to be trimodal, with comparable components as in Stone et al. Secondary/sensitivity analyses repeated the analysis for different baseline levels of depression severity, imputed values, and patient-reported depression symptoms.

Results: The best fitting models were either bimodal or trimodal but the trimodal solution did not meet criteria for replication. The bimodal model had one component with HDRS mean change of M = -13.0, SD = 6.7 and included 65.3% of patients, and another component with M = -1.8, SD = 5.1, 34.7%, respectively. For the trimodal model, the component with the largest change (M = -14.3, SD = 6.4) applied to 52% of patients, which differed substantially from the large effect component in Stone et al. (M = -18.8, SD = 5.1) which applied to 7.2%. Secondary/sensitivity analyses arrived at similar conclusions and for patient-reported depression symptoms the best fitting models were unimodal or bimodal.

Conclusions: This analysis failed to identify the trimodal distribution of response reported in Stone et al. In addition to being difficult to operationalize for regulatory purposes, results from mixture modeling are not sufficiently reliable to replace the more robust approach of comparing mean differences in depression rating scale scores between treatment arms.

Key words: antidepressants, biases, heterogeneity, average, treatment effect, efficacy

What is new

Key findings

- This mixture modeling analysis of a large, real-world antidepressant trial could not replicate the trimodal response distribution observed in industry trials. The best fitting models were unimodal, bimodal, or trimodal, depending on the choice of outcome or imputation method.

What this adds to what is known related to methods research within the field of clinical epidemiology

- It has been proposed that the average effect for antidepressants is misleading as this does not account for heterogeneity of response. This was supported by the finding of three subdistributions of response with finite mixture modeling in a large dataset of industry trials.
- The trimodal response distribution does not seem to be a robust finding and may be prone to biases. Our study suggests that the number and nature of the subdistributions via mixture modeling depend on study populations, imputation of missing data, and using clinician versus self-rated depression symptom scales.

What is the implication, what should change now

- It is premature to dismiss the average effect as the best estimate of antidepressant efficacy.

1. Introduction

The efficacy of antidepressant drugs is usually judged based on statistical significance and the size of the average drug-placebo difference in depression symptom scale scores. Although the average drug-placebo difference is statistically significant, the magnitude is small - about 2 points on the Hamilton Depression Rating Scale (HDRS) [1] which is below nearly all established thresholds of clinically meaningful effects [2,3].

However, efficacy judged by the average drug-placebo difference may be misleading if there is heterogeneity of treatment effects. In one study, outcome distributions which were non-normal and differed between drug-arms and placebo-arms were were statistically decomposed into two groups referred to as “non-benefiters” and “benefiters” and more patients in the drug-arm than in the placebo-arm were categorized as “benefiters” [4]. In a much larger recent analysis of individual patient data, the distributions of pre-post changes of depression symptom scores were decomposed into a mixture of different normal sub-distributions (modes, components) [1]. It was reported that a trimodal distribution best fit the data overall, which was interpreted to correspond to different subgroups of responders. The three sub-distributions corresponded with different levels of improvement, respectively denoted as “large” (with a mean change of M = 16.00 points, standard deviation SD = 4.22), “non-specific” (M = 8.94, SD = 6.96) and “minimal” (M = 1.68, SD = 2.99) response subgroups (Figure S1). 25% of antidepressant-treated patients were estimated to belong to the sub-distribution of “large” response, compared to 10% of those taking placebo. Stone et al. concluded that the small average drug-placebo differences are “best understood as affecting a minority of patients as either an increase in the likelihood of a Large response or a decrease in the likelihood of a Minimal response” (p. 5). While the term “response” is technically appropriate and commonly used, it is problematic as this may suggest that causal processes related to the treatment itself are involved in producing the subdistributions. However, besides the effect of treatment, several other mechanisms are involved in response, such as regression to the mean, natural course or methodological biases. Unfortunately, the findings have been interpreted as showing that there are specific subgroups of participants with distinct ‘responses’, which is misleading since the subdistributions overlap and do not correspond to groups of participants (see examples in the appendix). In addition, it remains unknown if the trimodal distribution is a robust finding. If such a similar distribution can be found in an open label, non-industry study, then this would be compatible with the assumption of a subgroup of patients with a large effect with whatever cause (actual drug effects, natural course, biases, both). If no trimodal distribution is found, especially if there is no subdistribution of a “large effect” then this raises questions about the external validity and interpretation of Stone et al.’s findings. Therefore, we wanted to explore whether Stone et al.’s finding of a trimodal distribution could be replicated by applying finite mixture modeling to the STAR*D study, a large single arm, open label, non-industry antidepressant trial

2. Methods

We conducted a secondary analysis of the STAR*D study (level 1). The analysis plan was registered on 2022-11-03 on the Open Science Framework (https://osf.io/rmdu9/) with the protocol uploaded prior to analysis. We used STROBE [5] as a reporting guideline.

2.1 Data

STAR*D is a large publicly funded study [6]. Enrolled patients were 18–75 years of age, seeking care at 18 primary and 23 psychiatric care clinics. Clinical research coordinators screened 4790 patients for major depressive disorder and administered the HDRS, on which 4041 patients scored ≥14, met the other inclusion criteria, and enrolled into the study. In level 1 of the study, all participants were treated open-label with citalopram for up to 14 weeks. HDRS-scores at the end of level 1 were obtained by independent, telephone-based interviewers. The Quick Inventory of Depression Symptomatology (QIDS-SR) self-report rating scale was regularly provided on site. In our main analysis, we selected the 3110 patients who scored ≥14 on the baseline HDRS because such cut-offs are common in clinical trials, too. In secondary analyses, we included patients independent of their baseline severity. We used a version of the STAR*D data as accessed through the NIMH Data Archive (collection ID #2148) in November 2019 by EP.

2.2. Analysis

2.2.1 Primary analysis

The primary analysis aimed to replicate the trimodal distribution of pre-post HDRS differences. We used a finite mixture modeling approach where non-normal distributions are decomposed into a set of different normal distributions, similar to Stone et al. [1]. We considered the replication successful if four prespecified criteria were met: a) the best fitting finite mixture model had three components (is trimodal), b) the order of the size of the components is comparable with Stone et al.'s findings among antidepressant treated patients (i.e., the number of “large” responders is smaller than the number of “non-specific” responders, and the number of “non-specific” and “minimal” responders differ minimally), c) the proportions of patients in each of the three components is comparable to those found in Stone et al., and d) the mean improvement in the “large” responder group is comparable to that found in Stone et al.

In their publication, Stone et al. used a finite mixture model with the data from the drug and placebo groups where means and standard deviations for the drug and placebo groups had to be identical (Figure S1). Because there was no placebo group in the STAR*D trial, we based the comparison on the results provided by Stone for antidepressant-treated patients for a modeling approach where means and standard deviations could vary for the drug and the placebo groups (Figure S1) .

We compared the proportions of patients in each of the components found between our study and Stone et al.’s using χ^2-tests, based on the 2 x 2 table (Study: Stone vs. STAR*D, category: large response vs. combined unspecific/minimal) and calculated effect sizes. The findings were considered comparable if the upper-bound of the confidence interval of the effect size did not overlap with Cohen’s $d = 0.3$. Similarly, the pre-post-differences within the large component were considered comparable if the upper-bound of the effect size did not exceed 0.3.

2.2.2. Secondary Analysis

We ran subgroup analyses with all patients independent of their baseline-severity which were categorized into baseline severity HDRS ≤ 18, 18 < HDRS ≤ 22, and HDRS > 22.

We also visually compared the distributions of the pre-to-post HDRS scores to those found in Stone et al.

2.2.3 Sensitivity Analyses

We ran two sensitivity analyses to see if the findings were sensitive to handling of missing data and choice of the outcome. First, we repeated the main analysis with imputed values for missing HDRS exit scores. Second, we repeated the main analysis with the patient self-report QIDS-SR as the outcome measure. See changes to the protocol for details below (2.4).

2.3. Statistical analysis

We used R 4.3.0 [7] for all analyses and the “mixR” package [8] to apply finite mixture modeling. We evaluated model fit using the Bayesian Information Criterion (BIC) and Akaike’s Information Criterion (AIC). Effect sizes were calculated using the “esc” package. Imputation was done using the “areg”-function of the “Hmisc” package [9]. Data was first accessed/analyzed between December 2022 and June 2024 (including a break due to lacking resources). Data was curated/provided by TK and CX and data analysis was performed by MP. Changes to the protocol are detailed below (2.4). The R-code is available via the Open Science Framework (https://osf.io/rmdu9/).

2.4. Changes to the initial protocol

For the primary analysis, we had originally planned to use the last observation carried forward (LOCF) approach to impute missing values for patients with missing exit HDRS scores. During data analysis we realized that there were no intermediate assessments of the HDRS and a LOCF approach would lead to an excess of pre-to-post differences of zero, thus obfuscating the finite mixture modeling approach. Therefore, for the primary analysis, we analyzed only those with available exit HDRS scores.

For the secondary analysis, we planned to impute missing exit HDRS values with multiple imputation using variables with less than 90% missing, but without providing further details. We decided (May 2024) to use multiple imputation where, for each imputed sample, “a flexible additive model is fitted on a sample with replacement from the original data and this model is used to predict all of the original missing and nonmissing values for the target variable" [9]. We generated 30 imputed samples, following the recommendation of Harrell (2015) [10]. For imputation we used the baseline demographic and clinical variables (Table 1) and the last available QIDS-SR values because these correlated highly with the HDRS ($r = 0.81$).

In the protocol we did not prespecify how to calculate Cohen's d for the $\chi2$-tests. Because of the theoretical priority to compare the components with the largest pre-post change in HDRS values, we decided (March 2024) to base it on the 2 × 2 table Study (Stone vs. STAR*D) × category (large response vs. combined unspecific/minimal).

3. Results

3.1. Study populations

A flow chart of the patient selection process is provided in Figure 1. Baseline characteristics of patients are provided in Table 1.

3.2. Primary Analysis

In the primary analysis using only complete cases, the best fitting models were bimodal according to the BIC and trimodal according to the AIC (Figure 2). However, the trimodal solution did not meet the other three predefined criteria for replication. First, we found a larger proportion of patients in the large response component than the non-specific response component (52.0% vs. 3.2%), whereas Stone et al. found the opposite result (7.2% vs. 41.8%). Second, the proportions in the three components were different in our study compared to those in Stone et al.'s. In our study, the proportion in the large response vs. all other components was 52% vs. 48%, compared to 7.2% vs. 92.8% in

Stone et al., resulting in a large difference, χ^2(df=1) = 5052.5, p < 0.01, d = -1.45 (95% CI -1.40 – -1.50). The proportions in the non-specific and minimal component were about equal in Stone et al. but differed substantially in our study, χ^2(df=1) = 609.7, p < 0.01, d = 1.34 (95% CI -1.21 – -1.48). Third, the pre-post improvement in the large-response component was -14.3 (95% CI -13.93 – -14.67) in our study and -18.8 (95% CI -18.63 – -18.97) in Stone et al., d = -0.83 (95% CI -0.76 – -0.90). The large response component in our study was more similar to the non-specific component in Stone et al. (M = -14.3 vs M = -14.8).

3.3. Secondary Analysis

Results by baseline severity

For the subgroup of patients with a baseline severity HDRS score ≤ 18, the best fitting model had two and three components according to the BIC and AIC, respectively (Figure S2). However, the components in the trimodal solution differed from those in Stone et al. because the proportions in the small and nonspecific response components differed substantially: 27.3% vs. 53.1% in the STAR*D study compared with 51.0% vs. 41.8% in Stone et al.

For the subgroup of patients with a baseline severity 18 < HDRS ≤ 22, the best fitting model had two components (Figure S3).

For the subgroup of patients with a baseline severity HDRS score > 23, the best fitting models had two and one component(s) according to the BIC and AIC, respectively (Fig. S4). No trimodal solution could be found because the model did not converge.

Visual comparison of distributions of pre and post HDRS scores

The baseline HDRS scores of the dataset in Stone et al. (their eFigure 2) seemed to be non-normally distributed with a tighter distribution centered at an HDRS score of 23. In contrast, the baseline HDRS scores in the STAR*D study seemed to be more normally distributed (Figure S8). We could not compare the posttreatment HDRS scores in the sample as a whole since this information was not available in Stone et al.

3.4. Sensitivity Analyses

In the sensitivity analyses where missing exit HDRS scores were imputed, the models converged in 28 of the 30 samples. The best fitting models according to the BIC were unimodal in 3 of the 28 imputed samples, bimodal in 24 samples, and trimodal in 1 sample (Figure S5). According to the AIC, the best fitting model was bimodal in 9 of the 28 imputed samples and trimodal 19 times. The trimodal solutions varied substantially in their nature, meaning that very different solutions fitted the data equally well (online supplement).

For the QIDS-SR as outcome, the best fitting model had one and two components according to the BIC and AIC, respectively (Figure S7).

4. Discussion

This study analyzed the HDRS pre-post differences in level 1 of the STAR*D study, where all patients were treated with citalopram, to see if the distribution is non-normal and better explained by sub-distributions similar to those in industry-sponsored clinical trials [1]. Using finite mixture modeling, Stone et al. [1] found that the non-normal distribution in their dataset was better explained by a trimodal distribution than a unimodal one, including a small component with a large mean change from baseline. In contrast, the best fitting model in the STAR*D data was either bimodal or trimodal, but the trimodal model differed substantially from Stone et al.'s so that none of the pre-specified criteria for replication was met. In particular, we did not identify a response component which was comparable with the large response component in Stone et al. We observed similar discrepant findings in secondary analyses for different baseline-levels of depression or for imputed values for the HDRS. In the sensitivity analysis with the self-report QIDS-SR measure, the best fitting model had only one or two components. Thus, none of the results from our analysis of the STAR*D data were in line with the findings by Stone et al.

Our results are relevant for the interpretation of findings from clinical trials, where the average efficacy of antidepressants is small and likely not clinically significant [2,3]. However, the average drug-placebo difference might be misleading if there is heterogeneity in treatment effects. This was suggested by the results of Stone et al.'s modeling analysis that the outcome distribution is non-normal and better explained by three response components. Patients classified into the "large" response component were suggested to be "(endo)phenotypes that are specifically responsive to antidepressant drugs (p. 5)". In our study, we could not replicate these findings, that is, finite mixture modeling results did not show that the trimodal model was consistently the best fitting model and there was no comparable distribution of "large" response.

If some patients respond especially well, as suggested by the trimodal distribution in Stone et al., then comparable distributions should also be seen in real-world trials such as STAR*D and not only in clinician rating scales but also in self-report scales. The failure to replicate the trimodal distribution in the STAR*D study and the different findings for clinician vs. self-reports raises doubts about the generalizability of the trimodal findings in RCTs and the finding that there is a subgroup of patients who respond especially well.

How could the discrepant findings be explained? For example, unblinding is present in most trials in which blinding is tested and this was associated with increased efficacy in some studies [11,12] but not in others [13]. Unblinding may lead to biased symptom ratings where improvement in the drug-arm is overestimated and improvement in the placebo arm is underestimated, leading to a shift of distributions. A tendency of clinicians to harmonize symptom ratings may explain bimodal response distributions, leading to overall scores clustering at either end of the distribution. A trimodal distribution could be explained if the rating bias interacts with degree of symptom-reduction, that is, if overestimation of improvement might be stronger for larger symptom-reductions, and underestimation of improvement might be stronger for smaller symptom-reductions. There is some evidence that improvement of symptoms on the HDRS towards symptom remission (e.g., from 1 to 0 on an HDRS item) is judged as being more important than other changes (e.g., from 3 to 2) [14,15]. If there are several HDRS items rated as zero, then this may bias ratings of other items more strongly towards improvement, compared to scenarios with no or few zero HDRS-item ratings. The effect of these putative biases on outcome distributions could be tested with simulations.

Rating biases may be less pronounced in non-industry trials such as the STAR*D trial, perhaps explaining why we could not replicate the trimodal distribution in Stone et al. Another finding raising doubt on the robustness of the trimodal solution is that it could not be found for self-report symptom ratings. Here, finite mixture models supported unimodal or bimodal response distributions. It would be interesting to repeat the analysis using self-report outcomes in industry trials. Furthermore, the distributions of drug and placebo arms should be more similar with more successful blinding. Other explanations for the differences between our results and those by Stone et al. are different recruitment and inclusion/exclusion criteria in industry trials and the STAR*D study, the variety of different antidepressants in Stone et al.'s study, or the use of different strategies to handle missing data. Comparison is also limited because there was no placebo control group in the

STAR*D study.

Our results suggest that further research is needed on the distributions of symptom measures in antidepressants trials and consideration of its implications. What constitutes a major deviation from the normal distribution and their difference between drug and placebo has not been discussed adequately yet, to our knowledge. The finite mixture modeling approach may deflect from the small average drug-placebo difference in antidepressant trials. If taken to its logical conclusion then even treatments with zero (or even negative) mean drug-placebo differences cannot be dismissed until sub-groups with large improvements can be ruled out, invalidating well-established testing paradigms for treatments. Furthermore, there is good reason to remain skeptical about the outlook to identify patients who benefit especially well from treatment [15] and until subgroups of patients who respond more or less to treatment cannot be predicted, results from statistical models to decompose the outcome are just descriptions of data. Unfortunately, the subgroups identified via finite mixture models are easily misinterpreted in at least two ways (examples in the Supplement). First, in the interpretation of Stone et al.'s study, the overlap between the subdistribution has been ignored and thus the drug-placebo differences in "response" have been overestimated. Second, the subgroups have been interpreted as distinct groups of patients caused by different effects of the treatment but this cannot be inferred from results of statistical models. Finally, reliably identifying subgroups of responders via finite mixture modeling requires large samples and should be seen in patient-reported outcomes, too. For smaller samples, qualitatively different solutions may have comparable fit and the results may be susceptible to imputation methods.

5. Conclusions

In conclusion, the trimodal antidepressant response distribution as reported in Stone et al. could not be replicated using data from the STAR*D trial, an open label, non-industry sponsored real-world antidepressant study. Therefore, our results do not support the notion that a subgroup of patients with a large response exists. Instead, these findings support the assumption that the putative subgroups from industry RCTs may be artifacts caused by methodological biases.

6. Additional Information

6.1. Acknowledgments

We thank Marc Stone for providing additional analysis and helpful discussions.

6.2. Data Statement

Data are not publicly available but are available from the NIMH-supported National Database for Clinical Trials (NDCT)

6.4. Funding Source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Stone MB, Yaseen ZS, Miller BJ, Richardville K, Kalaria SN, Kirsch I. Response to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the US Food and Drug Administration: individual participant data analysis. BMJ 2022;378:e067606. https://doi.org/10.1136/bmj-2021-067606.

[2] Hengartner MP, Plöderl M. Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression. BMJ Evid Based Med 2021. https://doi.org/10.1136/bmjebm-2020-111600.

[3] Moncrieff J, Kirsch I. Efficacy of antidepressants in adults. BMJ 2005;331:155–7. https://doi.org/10.1136/bmj.331.7509.155.

[4] Thase ME, Larsen KG, Kennedy SH. Assessing the ‘true’ effect of active antidepressant therapy *v.* placebo in major depressive disorder: use of a mixture model. Br J Psychiatry 2011;199:501–7. https://doi.org/10.1192/bjp.bp.111.093336.

[5] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. International Journal of Surgery 2014;12:1495–9. https://doi.org/10.1016/j.ijsu.2014.07.013.

[6] National Institute of Mental Health. NIMH » Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study. NIMH » Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study n.d. https://www.nimh.nih.gov/funding/clinical-research/practical/stard (accessed September 25, 2022).

[7] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.

[8] Yu Y. mixR: Finite Mixture Modeling for Raw and Binned Data 2021.

[9] Harrell F. Hmisc: Harrell Miscellaneous. R package version 5.1-0 2023.

[10] Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-19425-7.

[11] Jureidini J, Moncrieff J, Klau J, Aboustate N, Raven M. Treatment guesses in the Treatment for Adolescents with Depression Study: Accuracy, unblinding and influence on outcomes. Aust N Z J Psychiatry 2024;58:355–64. https://doi.org/10.1177/00048674231218623.

[12] Scott AJ, Sharpe L, Colagiuri B. A systematic review and meta-analysis of the success of

blinding in antidepressant RCTs. Psychiatry Res 2022;307:114297. https://doi.org/10.1016/j.psychres.2021.114297.

[13] Lin Y-H, Sahker E, Shinohara K, Horinouchi N, Ito M, Lelliott M, et al. Assessment of blinding in randomized controlled trials of antidepressants for depressive disorders 2000–2020: A systematic review and meta-analysis. EClinicalMedicine 2022;50:101505. https://doi.org/10.1016/j.eclinm.2022.101505.

[14] Kim TT, Xu C, DeRubeis RJ. Mapping Female Patients' Judgments of Satisfaction to Hypothetical Changes in Depression Symptom Severity. Behavior Therapy 2022;53:392–9. https://doi.org/10.1016/j.beth.2021.10.003.

[15] Kim TT, Xu C, Derubeis RJ. Patients' judgments of the importance of treatment-induced reductions in symptoms of depression: The role of specific symptoms, magnitudes of change, and post-treatment levels. Psychotherapy Research 2022;32:404–13. https://doi.org/10.1080/10503307.2021.1938731.

[16] Harrell F. Statistical Thinking - The Burden of Demonstrating HTE 2019. https://www.fharrell.com/post/demohte/ (accessed August 29, 2024).

Table 1. Sociodemographic and Clinical Characteristics (primary analysis, N = 3110)

	M or N	SD or %	% missing
Demographic Features			
Age	41.00	13.03	0
Female	1996	64	0
Race			
black	327	11	0
hispanic/latino	402	13	0
white	1977	64	0
other	404	13	0
Education (years)	13.60	3.23	47
Monthly household income	2321.86	2978.19	7
Employment			
employed	1711	55	0
unemployed	1205	39	0
retired	170	5	0
Insurance			
private	1517	4917	0
public	580	19	0
none	1047	34	0
Marital status			
single	905	29	0
married/cohabiting	1287	41	0
separated/divorced	823	26	0
widowed	9241	3	0
Clinical Features			
First episode age < 18	1200	39	1
Recurrent depression	1940	67	7
Family history of depression	1694	55	2

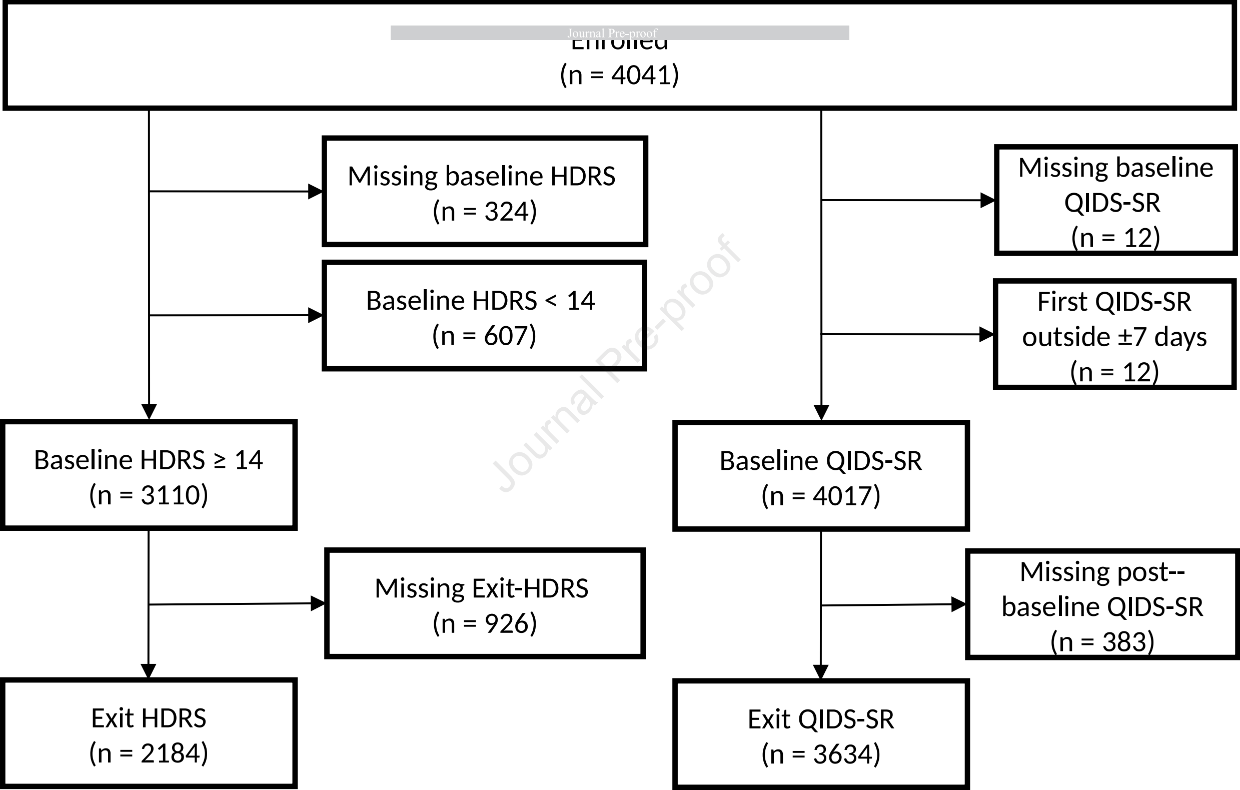
	M or N	SD or %	% missing
Age at first episode	25.14	14.29	2
Illness duration (years)	16.11	13.47	1
Number of episodes	5.56	9.32	15
Duration current episode (months)	24.88	52.02	1
Duration current episode >= 2y	787	26	1
QoL Questionnaire	39.07	14.26	12
SF-12 Mental	25.58	8.06	12
SF-12 Physical	48.61	12.13	12
Work and Social Adjustment Scale	24.98	8.67	12
HDRS-17	21.87	5.21	0
IDS-C30	39.07	9.64	2
QIDS-IVR	16.88	3.31	3
Cumulative Illness Rating Scale			
Categories endorsed	2.49	1.55	0
Total score	4.74	3.88	0
Severity score	1.83	0.81	10
Psychiatric Diagnosis Screening			
Agoraphobia	559	18	1
Alcohol abuse/dependency	371	12	1
Bulimia	607	20	1
Drug abuse/dependency	234	8	1
Generalized anxiety disorder	736	24	1
Hypochondriasis	336	11	1
OCD	723	23	1
Panic Disorder	422	14	1
PTSD	387	13	1
Social Anxiety disorder	963	31	1
Somatoform disorder	284	9	1
Number of axis I comobid psychiatric disorders	0.35	0.79	1

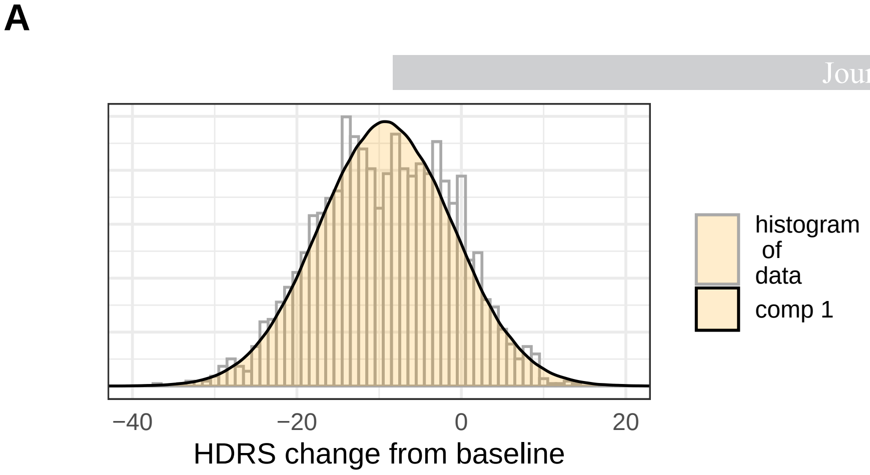
Note: Variables with less then 10% missing data were used for imputing the missing HDRS exit scores.

Figure Captions

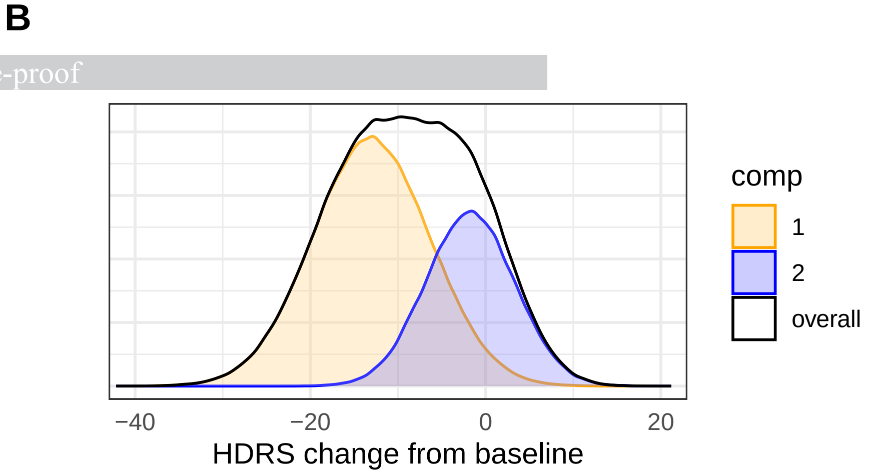
Figure 1. Patient flowchart.

Figure 2. Results of Finite Mixture Modeling with 1 to 3 components (panels A-C) and, for comparison, the results from the drug-arm in Stone et al (2022) for a model where the means and standard deviation could vary for both arms (panel D). The components are plotted in different colors. The distribution of the original values is plotted as a histogram in the background of panel A. The densities of the mixture models are plotted as thick black lines (“overall”).

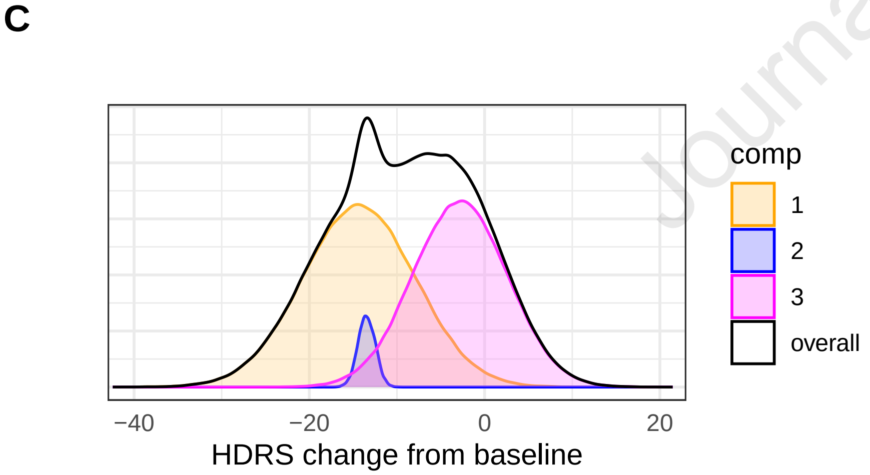

Enrolled
(n = 4041)
Missing baseline HDRS
(n = 324)
Baseline HDRS < 14
(n = 607)
Baseline HDRS ≥ 14
(n = 3110)
Missing Exit-HDRS
(n = 926)
Exit HDRS
(n = 2184)
Missing baseline QIDS-SR
(n = 12)
First QIDS-SR outside ±7 days
(n = 12)
Baseline QIDS-SR
(n = 4017)
Missing post--baseline QIDS-SR
(n = 383)
Exit QIDS-SR
(n = 3634)



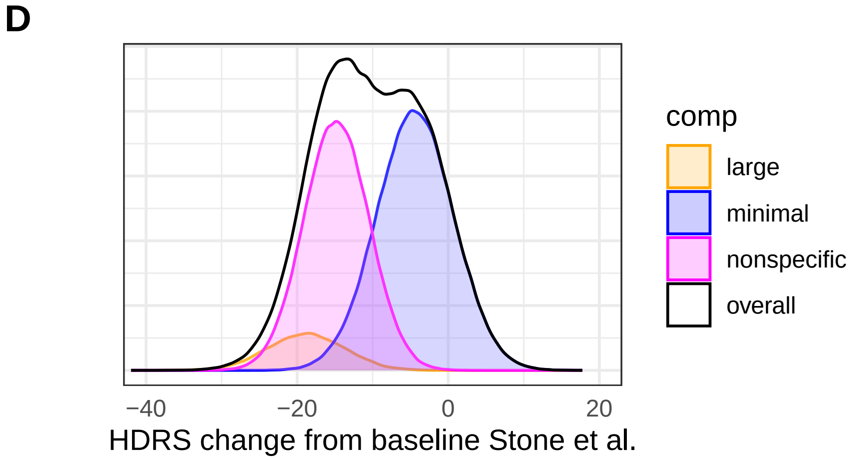
Model	Component	M (SD)	%	AIC	BIC
1	1	−9.13 (8.18)	100.00	15384.21	15395.59



Model	Component	M (SD)	%	AIC	BIC
2	1	−13.04 (6.70)	65.34	15351.08	15379.53
	2	−1.77 (5.07)	34.66		



Model	Component	M (SD)	%	AIC	BIC
3	1	−14.32 (6.41)	51.99	15348.56	15394.07
	2	−13.47 (1.00)	3.20		
	3	−2.80 (5.39)	44.81		



Model	Component	M (SD)	%
3	1	−18.8 (5.1)	7.2
	2	−14.8 (4.3)	41.8
	3	−4.4 (5.1)	51.0

JM receives royalties for three books about psychiatric drugs, and is a co-applicant on the REDUCE trial, funded by the National Institute of Health Research, evaluating digital support for patients stopping long-term antidepressant treatment. MAH and JM are both co-applicants on the RELEASE and RELEASE+ trials in Australia funded by the National Health and Medical Research Council (NHMRC) and Medical Research Future Fund (MRFF) evaluating hyperbolic tapering of antidepressants.MAH reports being a co-founder of Outro Health which aims to provide digital support for patients in the US to help stop no longer needed antidepressant treatment using gradual, hyperbolic tapering.

Florian Naudet received funding from the French National Research Agency (ANR-17-CE36-0010), the French ministry of health and the French ministry of research. He is a work-package leader in the OSIRIS project (Open Science to Increase Reproducibility in Science). The OSIRIS project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101094725. He is a work-package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. He is part of the Lown List of Industry-Independent Health Experts.

CX, EP, IK, MP, MPH and TK have no conflicts of interest to declare.