Use of Quantile Treatment Effects Analysis to Describe Antidepressant Response in Randomized Clinical Trials Submitted to the US Food and Drug Administration: A Secondary Analysis of Pooled Trial Data

William U. Meyerson, MD, PhD; Carl F. Pieper, DrPH; Rick H. Hoyle, PhD

Abstract

IMPORTANCE Major depressive disorder (MDD) is a leading cause of global distress and disability. Earlier studies have indicated that antidepressant therapy confers a modest reduction in depressive symptoms on average, but the distribution of this reduction requires more research.

OBJECTIVE To estimate the distribution of antidepressant response by depression severity.

DESIGN, SETTING, AND PARTICIPANTS In this secondary analysis of pooled trial data, quantile treatment effect (QTE) analysis was conducted from the US Food and Drug Administration (FDA) database of antidepressant monotherapy for patients with MDD, encompassing 232 positive and negative trials submitted to the FDA between 1979 and 2016. Analysis was restricted to participants with severe MDD (17-item Hamilton Rating Scale for Depression [HAMD-17] score ≥20). Data analysis was conducted from August 16, 2022, to April 16, 2023.

INTERVENTION Antidepressant monotherapy compared with placebo.

MAIN OUTCOMES AND MEASURES The distribution of percentage depression response was compared between the pooled treatment arm and pooled placebo arm. Percentage depression response was defined as 1 minus the ratio of final depression severity to baseline depression severity, expressed as a percentage. Depression severity was reported in HAMD-17-equivalent units.

RESULTS A total of 57 313 participants with severe depression were included in the analysis. There was no significant imbalance in baseline depression severity between the pooled treatment arm and pooled placebo arm, with a mean HAMD-17 difference of 0.037 points (P = .11 by Wilcoxon rank sum test). An interaction term test for rank similarity did not reject the rank similarity governing percentage depression response (P > .99). The entire distribution of depression response was more favorable in the pooled treatment arm than in the pooled placebo arm. The maximum separation between treatment and placebo occurred at the 55th quantile and corresponded to an absolute improvement in depression due to active drug of 13.5% (95% CI, 12.4%-14.4%). The separation between treatment and placebo diminished near the tails of the distribution.

CONCLUSIONS AND RELEVANCE In this QTE analysis of pooled clinical trial data from the FDA, antidepressants were found to confer a small reduction in depression severity that was broadly distributed across participants with severe depression. Alternatively, if the assumptions behind the QTE analysis are not met, then the data are also compatible with antidepressants eliciting more complete response in a smaller subset of participants than is suggested by this QTE analysis.
Introduction

Major depressive disorder (MDD) is a leading cause of global distress and disability. Antidepressants and psychotherapy are mainstays of MDD treatment. Most meta-analyses of antidepressant therapy randomized clinical trials focus on average treatment effect and/or its association with baseline characteristics. They generally find that patients with severe depression benefit from antidepressant therapy but only by a small amount on average, and this generates debate about whether use of antidepressants is worth the risk of adverse effects.

This debate also depends on how antidepressant efficacy is distributed in populations and in individuals, which is less commonly studied. Both population-level distributions and individual-level distributions matter, and the conceptual differences between them are subtle but important as explained in the eAppendix in Supplement 1.

Stone et al recently estimated the distribution of antidepressant efficacy with a high-quality US Food and Drug Administration (FDA) data set and inferred that 15% of participants experience a robust response specific to active drug. That study used mixture models to estimate the distribution of antidepressant response. This study by Stone et al is an important contribution to the literature, but we do not share the statistical assumptions of that work. Specifically, their model assumes that the mixtures identify the 3 natural, distinct subtypes of MDD patient response, whereas our belief is that the identified mixtures may represent statistical artifact. We wrote to Stone et al for access to a shareable portion of their data to reestimate the population and individual antidepressant efficacy distributions using different assumptions.

We model the distribution of antidepressant response using a quantile treatment effect (QTE) framework. When estimating the distribution of individual antidepressant response, we assume rank similarity. Rank similarity is a popular premise in QTE analysis that allows extrapolation from the population-level distribution of response to the individual-level distribution of response. Under the rank similarity premise, the expected counterfactual placebo response at a given quantile of response among participants in the drug arm is modeled as the actual placebo response at that quantile. For example, it assumes that the 55th percentile of individuals who respond to drug therapy, had they instead been assigned to placebo, would have experienced the depression response of the 55th percentile of placebo-assigned participants and likewise for other quantiles.

Rank similarity is appropriate when the features that affect a participant’s rank within one treatment arm exert a similar effect in both arms. This is clinically plausible in the depression context because many of the same factors are highly clinically significant in determining the course of a patient’s depression in either arm of a trial. For instance, patients with worsening social circumstances tend to have worse depression responses compared with their same-arm peers in the trial whether assigned to drug or placebo. Similarly, patients whose depression before the trial had been long-standing tend to have worse depression responses compared with their same-arm peers whether assigned to drug or placebo. Rank similarity would be violated if factors that affect antidepressant responsiveness were more important for determining relative rank within an arm than features that affect the course of depression regardless of treatment assignment. A recent review stated that such factors have not been conclusively identified, although we note that this is an area in which developments in precision medicine are still emerging. If rank similarity is not met, then QTE analysis that assumes rank similarity will tend to underestimate the amount of heterogeneity in treatment effects. For interested readers unfamiliar with QTE analysis or who seek more intuition regarding the rank similarity premise in the present context, we include a miniature, fully explained example in the eAppendix, eTable 1, and eTable 2 in Supplement 1.
Methods

Data Acquisition
This quantile treatment effects study is a secondary analysis of pooled participant data. We obtained aggregate data through personal correspondence with the authors of Stone et al., who had analyzed individual participant data (IPD) of 73,388 participants of the 232 positive and negative randomized clinical trials of antidepressant monotherapy for MDD disorder submitted to the FDA between 1979 and 2016. Data analysis was conducted from August 16, 2022, to April 16, 2023. Because the underlying individual studies and certain aspects of the IPD are proprietary, there were significant limitations on what form of data Stone et al. could share with us. Ultimately, we requested and received the following aggregate data for each combination of baseline depression severity and final depression severity: the number of individuals represented among their IPD in the treatment arm of any study and the number in the placebo arm of any study. These aggregate data imply a certain core set of IPD, such that we could extract for each participant their baseline depression severity, final depression severity, and whether they were assigned to the drug or placebo arms. In the aggregate data we received, depression severity had been converted to 17-item Hamilton Rating Scale for Depression (HAM-D-17) equivalents as reported by Stone et al., rounded to the nearest integer. We did not receive any study-level information or any other data about participants. This was the minimum necessary information for completing our planned QTE analysis. Because we used aggregated, anonymized data, our study was determined exempt from institutional review board approval by Duke University Health System. The analysis plan was not preregistered and no analyses were prespecified. We followed the relevant portions of the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline.

Statistical Analysis

Data Quality Testing and Processing
To test data completeness, we compared the counts of treatment-assigned and placebo-assigned participants in the aggregate data we received against the values reported by Stone et al. To test data consistency, we compared the range of baseline and final depression severity scores against the 0 to 52 range of the HAMD-17 scale. To test for risk of bias across studies introduced by the pooling process, we used a Wilcoxon rank sum test to compare baseline depression severity scores between treatment groups. When this test raised concern for a slight baseline imbalance, we used a literature-inspired filtering procedure of participants by baseline depression severity. A systematic review found that 20 is the most common HAMD-17 cutoff score for antidepressant trial inclusion among trials using the HAMD-17 scale; thus, we filtered out participants with baseline HAMD-17 scores less than 20. Then we again tested for baseline balance on the postfiltration set.

From the data obtained, we calculated 2 candidate measures of depression response since both percentage and absolute depression response are commonly used in studies. Percentage depression response was defined as 1 – F/B, expressed as a percentage, where F is the final depression severity and B is the baseline depression severity. Absolute depression response was defined as B – F. We used tests of rank similarity to guide which candidate measure of depression response to use in later parts of the study.

Testing Rank Similarity
We separately tested for rank similarity using our 2 candidate measures of depression response. We used the rank similarity test of Frandsen and Lefgren in which an available baseline attribute is tested for a significant interaction with the treatment arm when predicting the response, with the response separately tested as percentage depression response and absolute depression response. The available baseline attribute we used was baseline depression severity. Based on the results of these tests, we used percentage depression response as our chosen measure of depression response.
moving forward. A sensitivity analysis based on absolute depression response is presented in the eAppendix and eFigure in Supplement 1.

Quantile Treatment Effects
We calculated each quantile of percentage depression response separately in the treatment and placebo arms, at quantiles from the 5th percentile to 95th percentile, in increments of 5%. The QTEs were calculated as the difference between percentage depression response in the treatment vs placebo arms at a given quantile. The QTE 95% CIs were calculated by bootstrapping, using 10,000 iterations of the basic algorithm from the ci,quantile, diff function of the Hmisc R package, version 4.7-1. The QTEs were graphically plotted without covariates using the ci,qtet function of the qte R package, version 1.3.1 (R Foundation for Statistical Computing). Significance testing was performed via paired, 2-sided tests, with a significance threshold of \( P = .05 \).

Results

Data Quality
The aggregate data we received implied 71,393 participants, of whom 47,243 were assigned to drug and 24,150 to placebo. These values match the counts reported by Stone et al.\(^{15}\) All participants had an integer score for baseline depression severity and for final depression severity, and their range from 0 to 50 was within the 0 to 52 range of the HAMD-17 scale.

There was a slight and statistically significant difference in mean baseline depression severity between drug and placebo arms equivalent to 0.15 points on the HAMD-17 scale (\( P = 3.6 \times 10^{-5} \) by Wilcoxon rank sum test); this difference in baselines could have arisen from variable randomization ratios among the individual studies composing our pooled data. To address this difference, we noted that participants with very low HAMD-17 scores may not meet the criteria for MDD and that many published randomized clinical trials of antidepressants only include participants with a HAMD-17 score of 20 or greater.\(^{25}\) Thus, we excluded 9,115 drug-assigned participants and 4,965 placebo-assigned participants with baseline HAMD-17 scores less than 20 from our pool, yielding a final analysis set of 57,313 participants. After this filtration, the difference in mean baseline depression severity between drug and placebo arms decreased to 0.037 HAMD-17 points and was no longer significant (\( P = .11 \) by Wilcoxon rank sum test). That is, the synthesis procedure used to produce our pooled data had some evidence of bias, which we were able to mitigate through filtering. Due to our study design and limitations in available data, some additional common tests were not applicable; the eAppendix in Supplement 1 provides details.

Testing Rank Similarity
Since QTE analysis is most richly interpretable when rank similarity can be assumed, we tested whether our data and intended formulation were consistent with rank similarity. Following the method of Frandsen and Lefgren,\(^{19}\) we trained a linear model to predict percentage depression response from baseline depression severity, treatment arm, and the interaction between baseline depression severity and treatment arm. If there was a statistically significant interaction between baseline depression severity and treatment arm for prediction of percentage depression response in this model, then rank similarity would be rejected. When we applied this test, rank similarity was not rejected. Specifically, the interaction term between baseline depression severity and treatment arm in the model was not significant (\( P > .99 \)). While this test cannot prove that rank similarity holds, the results we obtained from this test provide some statistical reassurance in the plausibility of the rank similarity assumption for our context.

Since there is not consensus in the literature about whether to define depression response as a percentage change from baseline as we have done vs an absolute change from baseline, we also tested for rank similarity using absolute change from baseline as the response variable. Under this alternative formulation, the interaction term between baseline depression severity and treatment
arm in the model becomes significant ($P = .003$) (the eAppendix in Supplement 1 provides a
discussion of evidence that the magnitude of this interaction is very small). At a minimum this finding
indicates that a QTE analysis in which depression response is defined as an absolute change in
baseline must adjust for baseline depression severity. We instead choose to use percentage
depression response and not adjust for baseline depression severity.

**Quantile Treatment Effects**

Next, we characterized the estimated distribution of antidepressant response. We calculated the
depression response distribution separately in treatment and control conditions (Figure 1), then
Calculated QTEs as the difference between these distributions (Figure 2). We observed that
depression responses were more favorable in the treatment arm than in the placebo arm at all
reported quantiles. At the 55th quantile, treatment arm participants had a final depression score that
was 52.0% improved from baseline, and the corresponding value for placebo was 38.5% for a QTE
of 13.5% (95% CI, 12.4%-14.4%), with values at other quantiles listed in the Table. The QTEs were
greater in magnitude toward the center of the distribution and dissipated toward the tails. These

![Figure 1. Distribution of Percentage Depression Response](image1)

The distribution of percentage change of depression severity from baseline to trial end is depicted for each treatment arm. Error bars represent 95% CIs.

![Figure 2. Quantile Treatment Effects of Antidepressants Compared With Placebo](image2)

For each quantile ($\tau$) of depression response, the
difference between the treatment arm and placebo
arm depression response at that quantile is shown,
expressed as a percentage, along with its
bootstrapped 95% CI.
results suggest that, if rank similarity holds, then participants at any quantile of depression response experienced at least some additional response from antidepressant treatment.

**Discussion**

We conducted a secondary analysis of pooled IPD from 232 randomized clinical trials submitted to the FDA for antidepressant monotherapy for MDD to characterize the distribution of antidepressant response by depression severity. We observed that the population distribution of depression responses was strictly superior in the treatment arm vs the placebo arm in this filtered data set of participants with severe baseline depression. All studied portions of the depression response distribution were at least as favorable in the treatment arm as in the placebo arm. The participants with the worst response receiving antidepressant treatment responded at least as favorably (in terms of percentage improvement in HAMD-17 scores) as the participants with the worst response receiving placebo. Likewise, the participants with the best response receiving antidepressant treatment responded at least as favorably as the participants with the best response receiving placebo, and similarly for everyone in between. The response was most pronounced in the center of the distribution. These observations do not depend on any special assumptions; they are a simple consequence of the raw quantile distributions of antidepressant treatment response.

We did not observe violations of rank similarity in our data when using the percentage definition of depression response. In contrast, when we defined depression response in terms of absolute improvement from baseline, there was a formal violation of rank similarity, although the magnitude of that observed violation was small. Our ability to test for rank similarity was limited. For instance, we only had the data to test for the potentially rank-distorting influence of a single potential moderator of treatment effect and not several other potential moderators that have been reported in the literature, such as brain network and perfusion patterns.21-23 If the rank similarity premise is indeed true, then our results are compatible with the possibility that all individuals with MDD may experience at least slightly better depression responses while receiving antidepressant therapy compared with placebo.

**Table. QTEs at Selected Quantiles**

<table>
<thead>
<tr>
<th>Quantile within arm</th>
<th>Depression response, %</th>
<th>QTE (95% CI), %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Placebo</td>
</tr>
<tr>
<td>0.95</td>
<td>91.3</td>
<td>88.5</td>
</tr>
<tr>
<td>0.90</td>
<td>85.2</td>
<td>80.8</td>
</tr>
<tr>
<td>0.85</td>
<td>80.0</td>
<td>73.9</td>
</tr>
<tr>
<td>0.80</td>
<td>75.0</td>
<td>68.0</td>
</tr>
<tr>
<td>0.75</td>
<td>70.8</td>
<td>61.9</td>
</tr>
<tr>
<td>0.70</td>
<td>66.7</td>
<td>56.0</td>
</tr>
<tr>
<td>0.65</td>
<td>61.9</td>
<td>50.0</td>
</tr>
<tr>
<td>0.60</td>
<td>57.1</td>
<td>44.0</td>
</tr>
<tr>
<td>0.55</td>
<td>52.0</td>
<td>38.5</td>
</tr>
<tr>
<td>0.50</td>
<td>46.7</td>
<td>33.3</td>
</tr>
<tr>
<td>0.45</td>
<td>40.9</td>
<td>28.0</td>
</tr>
<tr>
<td>0.40</td>
<td>35.0</td>
<td>22.7</td>
</tr>
<tr>
<td>0.35</td>
<td>29.6</td>
<td>18.5</td>
</tr>
<tr>
<td>0.30</td>
<td>23.8</td>
<td>14.3</td>
</tr>
<tr>
<td>0.25</td>
<td>18.2</td>
<td>10.0</td>
</tr>
<tr>
<td>0.20</td>
<td>13.0</td>
<td>5.0</td>
</tr>
<tr>
<td>0.15</td>
<td>7.1</td>
<td>0.0</td>
</tr>
<tr>
<td>0.10</td>
<td>0.0</td>
<td>-3.7</td>
</tr>
<tr>
<td>0.05</td>
<td>-5.0</td>
<td>-9.5</td>
</tr>
</tbody>
</table>

Abbreviation: QTE, quantile treatment effect.
How should we interpret the fact that the middle of the quantile distribution is where the drug and placebo curves separate the most? Rank similarity permits the following interpretation: participants who would have had a typical, middling response to placebo (i.e., a percentage change of HAMD-17 score from baseline to final response at or near the median of the distribution of placebo responses) stand to gain the most from active drug. In contrast, patients who would have had a robust response with placebo receive little additional benefit from active drug. Perhaps this is because those with a robust response to placebo need no further treatment. Similarly, nonresponders to placebo would have received little additional benefit from active drug. Perhaps this is because individuals who do not respond to placebo experience a particularly entrenched form of depression that also responds poorly to drugs.

These findings are exploratory and would need to be confirmed through specialized placebo run-in trials with a prolonged run-in duration (e.g., the 6 weeks of placebo tested herein) followed by randomization of all run-in period participants, in contrast to common placebo run-in practices. The prediction is that partial responders in the run-in period would experience a greater benefit from active drug than would run-in nonresponders and run-in robust responders. If confirmed, future randomized clinical trials of antidepressants might increase their sensitivity by only randomizing partial responders of an adequate run-in period rather than the more common practice of randomizing nonresponders of a potentially inadequate run-in period.

Our analysis does not directly contradict that of Stone et al because that earlier work did not specifically preclude a large percentage of patients receiving small benefits from antidepressants and because we used a different response measure (percentage depression response) and a somewhat different data set (we excluded participants with low baseline HAMD-17 scores). Nonetheless, insofar as our study highlights the compatibility of our data with the possibility that all patients with severe depression may experience at least some improvement with antidepressant therapy, we provide a different perspective and emphasis than the focus of Stone et al on their estimated 15% of participants with substantial improvement while receiving antidepressants.

Analyses based on rank similarity tend to estimate response distributions as the broadest possible values that are consistent with the data. If rank similarity is only partly or approximately true, then the true distribution of antidepressant response would be more concentrated than the responses estimated herein. It would take a very large violation of rank similarity to arrive at responses as concentrated as those reported by Stone et al. While we believe that our statistical methods are well suited to uncovering distributions of response, it may be that the truth is somewhere between our estimates and those of Stone et al when it comes to individual patients.

Strengths and Limitations

A strength of this study is its use of FDA data, which include both published and unpublished high-quality randomized clinical trials and their pooled IPD. Another strength is that the population-level findings do not depend on any special assumptions.

One limitation of the study is that individual-level findings depend on the rank similarity assumption, which is unproven. Other limitations include lack of associated data and therefore omitted analyses concerning adverse effects, long-term effects, demographic characteristic covariates, and study-specific information. The lack of study-specific information in particular means that we cannot account for between-study heterogeneity, which might otherwise affect the study-adjusted drug and placebo response curves and the magnitude of calculated QTEs. In keeping with other QTE analyses in the literature, we did not attempt an analysis of very finely resolved quantiles, such as the bottom 1% of the population.

Moreover, although we do not consider our choice to exclude patients with less severe depression from the analysis as a limitation to the internal validity of our study, this choice potentially limits the generalizability of our results to patients with mild depression. That exclusion further limits direct comparison of our results with those of Stone et al.
Conclusions

Among participants with baseline HAMD-17 scores of 20 or higher from 232 FDA trials for antidepressant monotherapy for MDD, all segments of the distribution of depression responses were more favorable with active treatment vs control. These findings suggest that if our statistical assumptions hold, then there is the possibility that nearly all patients with MDD with baseline HAMD-17 scores of 20 or greater experience at least some benefit from antidepressant therapy, although the magnitude of the response is more clinically meaningful in some patients than others. If our assumptions are not met, it is also possible that the same aggregate benefit is concentrated in substantially fewer patients. Regardless, estimating the percentage of patients who benefit from antidepressant therapy is a challenging task that depends on the statistical assumptions used.

ARTICLE INFORMATION

Accepted for Publication: April 22, 2023.
Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2023 Meyerson WU et al. JAMA Network Open.

Corresponding Author: William U. Meyerson, MD, PhD, Department of Psychiatry & Behavioral Sciences, Duke University Medical Center, 3643 N Roxboro St, Durham, NC 27704 (william.meyerson@duke.edu).

Author Affiliations: Department of Psychiatry & Behavioral Sciences, Duke University Medical Center, Durham, North Carolina (Meyerson); Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, Connecticut (Meyerson); Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina (Pieper); Center on Aging and Human Development, Duke University Medical Center, Durham, North Carolina (Pieper); Department of Psychology & Neuroscience, Duke University, Durham, North Carolina (Hoyle).

Author Contributions: Dr Meyerson had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Meyerson, Hoyle.

Acquisition, analysis, or interpretation of data: Meyerson, Pieper.

Drafting of the manuscript: Meyerson.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Meyerson, Pieper.

Supervision: Hoyle.

Conflict of Interest Disclosures: None reported.

Data Sharing Statement: See Supplement 2.

Additional Contributions: John Beyer, MD (Department of Psychiatry & Behavioral Sciences, Duke University Medical Center), contributed to helpful discussions. No financial compensation was provided.

REFERENCES


SUPPLEMENT 1.
eAppendix. Supplementary Information and QTE Example
eTable 1. Simplified Illustration of QTEs Recovering Latent Natural Impact When Rank Assumptions Satisfied
eTable 2. Simplified Illustration of QTEs Misidentifying Latent Natural Impact When Rank Assumptions Violated
eFigure. Quantile Treatment Effects Framed as Absolute Change From Baseline

SUPPLEMENT 2.
Data Sharing Statement