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

Marc B Stone,[1] Zimri S Yaseen,[1] Brian J Miller,[2] Kyle Richardville,[3] Shamir N Kalaria,[4] Irving Kirsch[5]

[1]Division of Psychiatry, Office of Neuroscience, Office of New Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

[2]Division of Hospital Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]Department of Medicine, Cleveland Clinic Foundation, Cleveland, OH, USA

[4]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

[5]Program in Placebo Studies, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Correspondence to: B J Miller brian@brianjmillermd.com (ORCID 0000-0003-3247-8845)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2022;378:e067606
http://dx.doi.org/10.1136/bmj-2021-067606

Accepted: 02 June 2022

ABSTRACT

OBJECTIVES
To characterize individual participant level response distributions to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the US Food and Drug Administration from 1979 to 2016.

DESIGN
Individual participant data analysis.

POPULATION
232 randomized, double blind, placebo controlled trials of drug monotherapy for major depressive disorder submitted by drug developers to the FDA between 1979 and 2016, comprising 73 388 adult and child participants meeting the inclusion criteria for efficacy studies on antidepressants.

MAIN OUTCOME MEASURES
Responses were converted to Hamilton Rating Scale for Depression (HAMD17) equivalent scores where other measures were used to assess efficacy. Multivariable analyses examined the effects of age, sex, baseline severity, and year of the study on improvements in depressive symptoms in the antidepressant and placebo groups. Response distributions were analyzed with finite mixture models.

RESULTS
The random effects mean difference between drug and placebo favored drug (1.75 points, 95% confidence interval 1.63 to 1.86). Differences between drug and placebo increased significantly (P<0.001) with greater baseline severity. After controlling for participant characteristics at baseline, no trends in treatment effect or placebo response over time were found. The best fitting model of response distributions was three normal distributions, with mean improvements from baseline to end of treatment of 16.0, 8.9, and 1.7 points. These distributions were designated Large, Non-specific, and Minimal responses, respectively. Participants who were treated with a drug were more likely to have a Large response (24.5% *v* 9.6%) and less likely to have a Minimal response (12.2% *v* 21.5%).

CONCLUSIONS
The trimodal response distributions suggests that about 15% of participants have a substantial antidepressant effect beyond a placebo effect in clinical trials, highlighting the need for predictors of meaningful responses specific to drug treatment.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Clinical trials of antidepressants in major depressive disorder show substantial mean improvement with both drug and placebo

Meta-analyses have confirmed that antidepressants have greater efficacy than placebo, but the mean difference is small

WHAT THIS STUDY ADDS

After accounting for participant baseline severity, age, and sex, placebo responses and drug effects were stable over time

Antidepressants and placebo showed the same three modal responses

The small mean advantage of antidepressants is because of differences between drug and placebo in a minority of participants in the likelihood of achieving a Large response or avoiding a Minimal response

Introduction

Depression is a leading cause of disability worldwide, affecting 300 million people globally, causing a major reduction in quality of life, with domestic costs (including costs related to work) estimated at more than $210.5 (£175.3; €207.1) billion annually.[1 2] About 13% of Americans use antidepressants, and use of antidepressants in economically developed countries more than doubled between 2000 and 2015.[3 4] Although many factors affect depression and suicide rates, the hope was that wider use of antidepressants would improve these rates. Nonetheless,[5] these rates have generally increased,[6] particularly in younger age groups, highlighting the importance of understanding the magnitude and determinants of the efficacy of antidepressant drugs.

Previous reviews have assessed the effects of antidepressants by analyzing aggregate trial data[7-14] or participant level data from limited datasets. Meta-analyses have shown small mean differences between drug and placebo arms, and the clinical significance of these differences continues to be debated.[7-18] Patients do not feel the difference in response between drug and placebo (drug effect); rather, patients have an overall drug response in the context of pharmacotherapy. How much was attributable to placebo effects is unobservable. In this paper, we use the term drug or placebo response to indicate change from baseline with the drug or placebo, and the term drug or placebo effect to indicate the component specifically attributable to the drug or placebo.[19]

Lack of knowledge about the distributions of individual responses has hampered discussions of the clinical significance of mean effects. Whether treatment responses in clinical drug trials are best described by one or multiple underlying distributions of treatment response or how drug and placebo response distributions differ is not known. The drug effect might not be a uniform small, and hence clinically unimportant, benefit across patients (ie, a shift in distribution mean without a change in the shape of the distribution). Rather, it could occur as a large, and thus clinically important, difference for a small subpopulation (ie, a difference in response distribution composition). Some investigators have attempted to look at this possibility by comparing variability in treatment response in patients treated with a drug or placebo. These analyses cannot rule out the effects of restricted subpopulations, however.[20-22] Researchers also continue to debate the relation between the initial severity of the disease and the effect of the drug,[14 23-25] which patient subgroups benefit most from antidepressant treatment,[26 27] and whether new trials are hampered by rising placebo response rates.[28 29]

In this article, we report a participant level analysis of randomized, placebo controlled trials of acute monotherapy for the treatment of major depressive disorder submitted to the US Food and Drug Administration from 1979 to 2016. We used mixture modeling of the distributions of participant level responses in randomized, placebo controlled trials of antidepressant drugs to determine subpopulations that the response distributions might comprise and to determine whether the difference between drug and placebo can be accounted for by a broadly applicable small incremental effect of the drug. We supported our aggregation of subject level trial data and looked at other controversies relating to efficacy trials of acute antidepressant treatment, with examinations of how mean responses and baseline severity have changed over time. We also looked at relations between age, sex, baseline severity of depression, and their interactions, and mean responses.

Methods

Our database contained 232 randomized, double blind, placebo controlled trials of drug monotherapy for major depressive disorder submitted by drug developers to the FDA, comprising 73 388 participants. The database included studies in new drug applications, whether positive or negative, and whether an indication for major depressive disorder was applied for or approved. The dataset also included studies after approval. All studies had specified study objectives and inclusion and exclusion criteria, as required to meet regulatory standards, thus meeting typical criteria for high quality studies and low risk of bias. Also, studies of clinical efficacy submitted to the FDA are generally reviewed by expert reviewers for fitness for purpose before the start of the study if conducted in the US, providing additional assurance of study quality. The data elements used for this study were trial and participant identifiers, treatment assignment (specific drug or placebo), age, sex, primary scale used to measure the severity of symptoms of major depressive disorder, and the score on that scale at baseline and at the last observation on treatment.

To conduct analyses of the severity of major depressive disorder across trials that used different instruments, we converted all scores to equivalent 17 item Hamilton Rating Scale for Depression (HAMD17) scores. HAMD17 was the most widely used measure (in 104 of 232 trials) in the dataset. The supplementary material has details of the conversions.

A mixed effects model with study as random effect estimated the mean difference in change from baseline (last measurement before treatment) for drug compared with placebo, as well as change from baseline for individual participants adjusted for random study effects. Other models included age, sex, baseline severity, and their interactions with treatment assignment, and with each other. Ecological bias in treatment covariate interactions was avoided by centering the participant level covariate around the study mean, as recommended by Burke et al.[30] Because baseline severity correlates artifactually with change from baseline but not with severity at the end of treatment, models of change from baseline that included baseline severity as a predictor, estimated end of treatment severity rather than change from baseline and were converted into estimates of change from baseline by subtracting baseline severity. Models that used date of study participation as an explanatory variable were used to look for trends over time. Multivariable adaptive regression splines, with the method of Royston and Sauerbrei,[31] were used to account for non-linear effects of continuous explanatory variables. To explore differences in efficacy among individual drugs, we modeled the residuals from the random effects model that included all the non-linearities and interactions among age, sex, baseline severity, and drug versus placebo assignment as a function of individual drug assignment.

We used finite mixture modeling to evaluate whether response distributions were compatible with drug effects having a broadly applicable incremental improvement over placebo or whether response distributions were instead consistent with a combination of simpler distributions, as might be expected with different response populations. The models tested allowed for different numbers of component distributions and differences between drug and placebo groups, both in component distribution means and in proportional contributions. We also compared models of normal distributions with models of left skewed or right skewed log normal distributions. Minimization of the Akaike and Bayesian information criteria was used to select the best model, along with a requirement that each latent distribution represented at least 1% of the population. We also applied finite mixture modeling to a split sample randomized by trial and to several subgroups: men and women; mild, moderate, and severe baseline severities according to the criterial of the National Institute for Health and Care Excellence (NICE); and participants aged <18 years to evaluate model consistency across subsamples. eTable 3 provides more details of mixture model testing and selection. All statistical analyses were performed with Stata versions 15.1, 16.1, and 17.0.

Patient and public involvement

Patients and members of the public were not involved in the design, conduct, reporting, or dissemination of the research. Patients were not involved in the study planning process because industry study data are non-public and submitted to the FDA as part of the agency's regulatory processes.

Results

Sample characteristics and relations among baseline variables

Table 1 shows sample characteristics of the population. The random effects mean age of participants was 41.8 years, with 90% of participants aged 15-70.4 years. Table 2 and table 3 summarize baseline severity of depression (last assessed before randomized treatment) by demographic characteristics. Baseline severity was considerably lower for participants aged <18 years, particularly those aged ≤12 years. This finding was mostly because a lower level of severity was seen in trials that included only children, particularly in trials that included children aged <12 years. When individuals aged 16 and 17 years were included in adult trials, their mean baseline severity was 20.3 points; in pediatric trials, their average severity was 17.7 points. In pediatric trials with a minimum age of 12 or 13 years, the average baseline severity was 18.8 points; for trials that included participants aged <12 years, the average baseline severity for participants aged ≥12 years was 17.3 points. Within trials, the distribution of baseline severity was slightly skewed, with median severity being about 0.3 points lower than mean severity (eFig 2).

Treatment effects

The random effects mean changes (supplement eTable 2) were improvements of 9.8 points (95% confidence interval 9.5 to 10.0) with active drug and 8.0 points (7.8 to 8.3) with placebo. The difference between drug and placebo was 1.75 points (1.63 to 1.86). The magnitude of the difference was unchanged when the analysis was done separately in subgroups with native HAMD17 scores (1.75, 95% confidence interval 1.57 to 1.93; standardized mean difference 0.232, 95% confidence interval 0.210 to 0.255) and converted scores (1.75, 1.59 to 1.91; 0.245, 0.223 to 0.267).

Influence of and interaction among participant characteristics

When sex was included in the model as the only covariate, little difference in response to placebo was found (0.14 points less improvement in men, 95% confidence interval −0.05 to 0.33, P>0.1). For active drug, the difference was greater (0.35 points less improvement in men, 0.21 to 0.49).

When baseline severity was used as the only covariate, improvement with drug and placebo increased with greater baseline severity (eFig 3). The advantage of drug over placebo increased with baseline severity by 0.09 points (95% confidence interval 0.06 to 0.12) for every one point increase in severity. The estimated difference between drug and placebo at a baseline severity of 16 points (5th centile) was 1.1 points, increasing to 2.5 points at a baseline severity of 29.6 points (95th centile).

With age as the only covariate, we found a linear relation between age and response to placebo for adults; improvement over baseline diminished by 0.30 (95% confidence interval 0.22 to 0.38) points for every decade increase in age. The observed response for adults to active drug also diminished with age.

When sex, baseline severity, and their interaction were included in the model, the differences between sexes in response to placebo (P=0.008) and active drug (P=0.02) were slight but statistically significant. Differences between drug and placebo increased similarly with greater baseline severity for both sexes.

When age, sex, and their interaction were included, both sexes showed a similar (P>0.4 for a difference) response to placebo that decreased with age, but with active drug the difference between sexes increased with age by an estimated 0.13 (95% confidence interval 0 to 0.25) points per decade. For women, the largest improvement over baseline with drug was estimated as 10.1 points at age 30 years; the largest difference between drug and placebo was estimated as 2.2 points at age 62 years. For men, the largest improvement over baseline with drug was estimated as 9.8 points between ages 22 and 23 years; the largest difference between drug and placebo was estimated as 1.7 points at age 57 years.

When age, baseline severity, and their interaction were included, these factors were strong predictors of change from baseline for both drug and placebo, and for difference between drug and placebo. For both drug and placebo, improvement from baseline increased with baseline severity and decreased with age. Comparing pediatric (age <18) participants directly with adults, the unadjusted difference between drug and placebo (eTable 2) was 1.12 points (95% confidence interval 0.66 to 1.57) greater for adults. Adjusted for baseline severity, the difference between drug and placebo was greater in adults by 0.64 points (0.17 to 1.12). Figure 1 shows the interactions among treatment, age, sex, and baseline severity. Generally, the greatest difference between drug and placebo was seen in older participants with higher baseline severity.

Trends over time in study characteristics and results

We found no detectable trend over time in sex distribution (P>0.15). The average age of participants was consistently 42 years until about 2005 when a notable downward trend was seen attributable to a relative absence of older adult participants, with an additional steep decline beginning in 2013 because most trials were conducted in children (79%

Table 1 | Population characteristics

	No (%)	Percentage range among trials or No of trials
Total	73 388	—
Sex:		
Men	28 738 (39.3)	9.0-77.6
Women	44 478 (60.7)	22.4-91.0
Age (years):		
Pediatric (6 to ‹18)	4896 (6.7)	0.0-100
6 to ≤12	1952 (2.7)	0.0-57.3
›12 to ‹18	2944 (4.0)	0.0-97.5
Adult (≥18)	68 492 (93.3)	0.0-100
≥18 to ›30	12 174 (16.6)	0.0-42.7
≥30 to ›50	33 110 (45.1)	0.0-73.3
≥50 to ›65	15 670 (21.4)	0.0-75.2
≥65	7538 (10.3)	0.0-100
≥75	2305 (3.1)	0.0-99.4
Treatment assignment:		
Placebo	24 711 (33.8)	11.2-68.0
Antidepressant	48 495 (66.2)	32.0-88.8
Amitriptyline	625 (0.9)	9
Bupropion	3179 (4.3)	21
Citalopram	2340 (3.2)	14
Clomipramine	132 (0.2)	1
Desipramine	315 (0.4)	6
Desvenlafaxine	3956 (5.4)	15
Dothiepin	106 (0.1)	1
Duloxetine	3865 (5.3)	24
Escitalopram	2145 (2.9)	14
Fluoxetine	5179 (7.1)	40
Fluvoxamine	1532 (2.1)	16
Imipramine	1908 (2.6)	23
Levomilnacipran	1579 (2.2)	5
Mirtazapine	1166 (1.6)	13
Nefazodone	2154 (2.9)	18
Paroxetine	4250 (5.8)	26
Selegiline	817 (1.1)	5
Sertraline	3206 (4.4)	24
Trazodone	121 (0.2)	2
Venlafaxine	3932 (5.4)	29
Vilazodone	2394 (3.3)	9
Vortioxetine	3594 (4.9)	12
Length of study (weeks):		
3	50 (0.1)	1
4	1470 (2.0)	13
5	517 (0.7)	5
6	16 964 (23.1)	62
7	930 (1.3)	4
8	37 740 (51.4)	99
9	1759 (2.4)	6
10	6389 (8.7)	16
11	105 (0.1)	1
12	5154 (7.0)	16
13	637 (0.9)	2
14	575 (0.8)	2
15	439 (0.6)	2
16	400 (0.5)	2
Unknown	259 (0.4)	1
Study size (No of participants):		
≤50	224 (0.3)	5
51-100	1285 (1.8)	18
101-200	7965 (10.9)	53
201-400	26 008 (35.4)	90
401-600	21 772 (30.0)	47
601-800	12 125 (16.5)	16
›800	4009 (5.5)	3
Year of study participation:		
Before 1995	7501 (10.2)	13
1995-99	18 136 (24.7)	45

(Continued)

of participants were aged <18 years). The random effects mean severity at baseline decreased by 1.54 (95% confidence interval 0.47 to 2.61) points, mostly between 1979 and 1995; a further reduction after 2013 was because most of the trials were conducted in children. End of treatment severity seemed to decrease slightly for active drug and placebo, but these trends were not statistically significant. After adjustment for age, sex, and baseline severity, no evidence for change in treatment responses over time was seen (P>0.7).

Response distributions

Figure 2 shows the distributions of responses to drug and placebo (compared with the superimposed modeled distributions). With drug, 41 790 (88.5%) of 47 243 participants showed some improvement (compared with 20 376 (84.4%) of 24 150 for placebo) and the median improvement was 9.8 points (compared with 7.2 points for placebo).

The distributions of responses for drug and placebo did not appear unimodal. Analysis with finite mixture modeling found that the optimal model for drug and placebo responses was a combination of three overlapping normal distributions allowed to vary in relative size between drug and placebo (fig 2). The respective modeled overall distributions differed only in the proportions drawn from the underlying latent distributions (corresponding to the area under the curve for each). Allowing different means or skewness in the latent distributions for drug and for placebo did not improve the fit of the model. This trimodal normal model was robust across random subgroups and subgroups defined by baseline characteristics, and was consistently one of the two best models by Akaike and Bayesian information criteria (eTable 3). When other models (including models with four modes) showed a better value for the Akaike or Bayesian information criterion (although never both), no similar consistency across subgroups was found; rather, they seemed to deal with minor deviations from normality in the data.

One latent distribution (Large) represented a large degree of improvement (mean improvement 16 points, standard deviation 4.2), one (Minimal) represented little or no improvement (1.7, 3.0), and the third (Non-specific) represented a broad range (8.9, 7.0). Compared with placebo, the distribution for active drug was more likely to show Large responses (24.5% *v* 9.6%, odds ratio 3.07, 95% confidence interval 2.05 to 4.91) and less likely to show Minimal responses (12.2% *v* 21.5%, 0.51, 0.41 to 0.62). Most responses (63.3% of active drug and 68.9% of placebo), however, were in the Non-specific category. The estimated number needed to treat with active drug to realize one more patient with Large improvement was 6.7 (95% confidence interval 5.7 to 7.7). The number needed to treat with active drug to realize one less patient with Minimal improvement was 10.8 (9.0 to 12.5).

Drug level differences in treatment effect

Figure 3 shows differences in effect size among active drugs, adjusted for age, sex, and baseline severity.

Table 1 | Continued

	No (%)	Percentage range among trials or No of trials
2000-04	20 087 (27.4)	87
2005-09	12 110 (16.5)	28
After 2009	8793 (12.0)	19
Unknown	6761 (9.2)	21
Study completers:*	45 132 (70.9)	212
Antidepressant	29 946 (70.3)	212
Placebo	15 186 (72.0)	212
Primary depression scale:		
HAMD17	30 393 (41.8)	104
HAMD21	15 588 (21.4)	59
HAMD24	2899 (4.0)	6
HAMD28	1737 (2.4)	5
HAMD31	724 (1.0)	2
MADRS	16 789 (23.1)	37
CDRS	3720 (5.1)	15
Other	937 (1.3)	4
HAMD17 score:		
Native	35 021 (47.7)	119
Converted	36 374 (49.6)	110
Change from baseline missing:	1993 (2.7)†	113
Antidepressant	1252 (2.6)	105
Placebo	561 (2.3)	93

HAMD=Hamilton Rating Scale for Depression, with number of items; MADRS=Montgomery-Asberg Depression Rating Scale; CDRS=Children's Depression Rating Scale.

*Data available at trial level, missing for 20 trials (9724 (13%) participants); denominator does not include missing data.

†180 participants who left the study before the start of treatment or assignment of treatment.

The drugs showing the largest beneficial effect (amitriptyline, clomipramine, and venlafaxine) also showed larger effects in trials where they were directly compared with other agents. These drugs represented about 10% (4689 of 48 495) of participants who received active drug, and the distribution of responses for these drugs (28% Large, 62% Non-specific, 10% Minimal) differed modestly from the other antidepressants (24% Large, 63% Non-specific, 12% Minimal).

Discussion

Principal findings and comparison with other studies

This participant level analysis of all placebo controlled monotherapy antidepressant efficacy trials submitted to the FDA between 1979 and 2016 provides a comprehensive participant level analysis of treatment responses and net drug effects. Consistent with meta-analyses, where the effects of antidepressant drugs on HAMD17 ranged from 1.62[24] to 2.56,[15] with standardized mean differences of 0.23[13] to 0.34,[32] we found a drug effect among adults equivalent to 1.82 points, with a standardized mean difference of 0.24. For pediatric participants, the drug effect was 0.71 points, with a standardized mean difference of 0.13. Some studies have attributed the relatively small standardized mean differences to increases in the placebo response over time[28 29 33] but this hypothesis was not supported in our study.

Table 2 | Depression severity at baseline—HAMD17 equivalent scores by demographic group

Study group	Random effects (mean)	Standard deviation (within study, total)	Range of participant scores across studies*	90% range	Random effects mean difference between groups† (95% CI), P value
All participants	22.6	3.5, 4.2	0-46	16-28	NA
Sex					
Women	22.7	3.5, 4.1	1-46	16-29.9	0.35 (0.30 to 0.41), <0.001
Men	22.3	3.4, 4.2	0-42	15.1-29	
Age (years)					
7 to ≤12 years	16.6	3.2, 3.4	3.3-31.5	11.1-21.9	0.75 (0.54 to 0.97), <0.001
>12 to <18 years	17.8	3.4, 3.9	5-37	12-24.5	
Adult age groups					
1 (18-33 years)	23.0	3.4, 3.9	2-40	17.2-29.2	Reference
2 (>33-43 years)	23.1	3.4, 4.0	0-46	17-29.8	0.07 (0.0 to 0.15), 0.053
3 (>43-54 years)	23.2	3.4, 4.0	2.6-44	17.2-30	0.11 (0.03 to 0.18), 0.004
4 (>54-100 years)	23.2	3.7, 4.3	1-46	17-30	0.14 (0.05 to 0.22), 0.002

NA=not applicable; HAMD17=Hamilton Rating Scale for Depression.

*Participants might not be depressed at baseline owing to improvement between screening and baseline.

†Difference in means might not equal random effects mean difference in mixed effects models.

Similar to Thase et al,[34] we found that a multimodal mixture model was the best fit for the data. However, we found that participants seemed to belong to one of three types of response populations. About two thirds of participants assigned drug and placebo had a Non-specific response. Those treated with drug were more likely to show a Large response (24.5% *v* 9.6% with placebo), however, and less likely to have a Minimal response (12.2% *v* 21.5%). Thus the observed advantage of antidepressants over placebo is 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. Examination of response distributions by demographic and baseline factors (eFigs 9-12) showed differences in the overall distributions between subgroups, consistent with our findings in the regression model of the effects of baseline severity and age on treatment responses. The subgroup distributions were best described by trimodal models (eTable 3), however, showing that our findings were not artifacts of response patterns of different subgroups. This result highlights continued potential for identification of (endo)phenotypes that are specifically responsive to antidepressant drugs.

The Non-specific distribution included a broad range of responses. Its distinctly broad range and similar likelihood for drug and placebo groups suggests that these responses might reflect the diverse interactions of individual characteristics with placebo and other effects not related to drug treatment, such as response to increased clinical contact, spontaneous improvements, and regression toward the mean. The Large and Minimal response distributions were distinguished from the Non-specific category by differences in means and smaller standard deviations. The Large response was more than twice as likely with drug than placebo and might be qualitatively different, showing a change in the depressive state rather than symptomatic attenuation; >90% of those having a Large response were subthreshold or not depressed according to the NICE criteria.

Implications

The trimodal model provides a different understanding of what is meant by clinical response. Response has been conventionally defined in the literature in clinical trials for depression by a 50% or greater improvement from

Table 3 | Depression severity at baseline across studies by HAMD17 equivalent score

Severity classification[32]	% of participants	Median % across studies	90% of % range across studies	% range across studies
APA/NICE not depressed (HAMD17 ≤7)	0.2	0	0-1.8	0-15.7
APA mild/NICE subthreshold (7 ‹HAMD17 ≤13)	1.7	0	0-15.7	0-24.9
APA moderate/NICE mild (13 ‹HAMD17 ≤18)	10.7	5.8	0-48.1	0-72.7
APA severe/NICE moderate (18 ‹HAMD17 ≤22)	35.9	36.3	10.3-54.0	0-65.3
APA very severe/NICE severe (22 ‹HAMD17)	51.5	52.2	5.4-88.0	0-100
CGI not at all to mildly ill (HAMD17 ≤15)	4.1	0	0-34.0	0-47.9
CGI moderately ill (15 ‹HAMD17 ≤22)	44.4	46.9	12.0-68.1	0-88.3
CGI markedly ill (22 ‹HAMD17 ≤28)	43.6	43.0	5.4-67.5	0-92.5
CGI severely to extremely ill (28 ‹HAMD17)	8.0	5.4	0-32.9	0-61.1

HAMD17=Hamilton Rating Scale for Depression; APA=American Psychiatric Association; NICE=National Institute for Health and Care Excellence; CGI=Clinical Global Impression.

baseline. These threshold definitions, although useful, are arbitrary. Also, comparing numbers of participants in response categories defined by thresholds cannot distinguish between a uniformly shifted distribution pushing a subset of patients over a threshold, and one where different subgroups have different treatment effects. Thus unlike latent distributions in a mixture model, outcome categories defined by thresholds also cannot suggest differential underlying processes for how a change in symptoms came about. The finite mixture models presented in this paper give insight into how participants might have achieved their responses. A useful analogy might be the difference between phenotype and genotype. The observed improvement is analogous to phenotype, whereas underlying response distributions are analogous to genotypes.

Our findings could therefore suggest a different framing of the discussion of the balance of potential benefits and harms for acute prescribing of antidepressant drugs. Our results suggest that rather than being likely to provide a small incremental benefit in reducing symptoms (beyond a placebo response), there is a modestly increased likelihood of providing a large near term reduction in symptoms or preventing continued near term symptom severity that would not have occurred otherwise. Also, patients with only a modest improvement in symptoms with an adequate trial of acute antidepressant treatment might not be having a drug specific effect (92% of those having a Large response improved by ≥10 points) and thus might need to switch to another treatment.

Previous analyses of the relation between baseline severity and the efficacy of antidepressants found null[23 24] to moderate (slope about 0.3)[14 16] effects, and these effects might be attributable to instrument behavior rather than patient experience.[25] We found that the effect of baseline severity was statistically significant (P<0.001) but small (slope about 0.1). This finding could be partially because of the ceiling effects on improvement among participants who were less depressed but might also be attributable to an increasing likelihood of a specifically drug responsive phenotype among those with greater symptom severity. In line with the NICE guidelines,[32] given the modest absolute likelihood of substantial benefit over placebo, and when consistent with availability and patient preferences, beginning with lower risk treatments for mild to moderate acute depression might be preferable if no underlying dysthymic disorder exists.

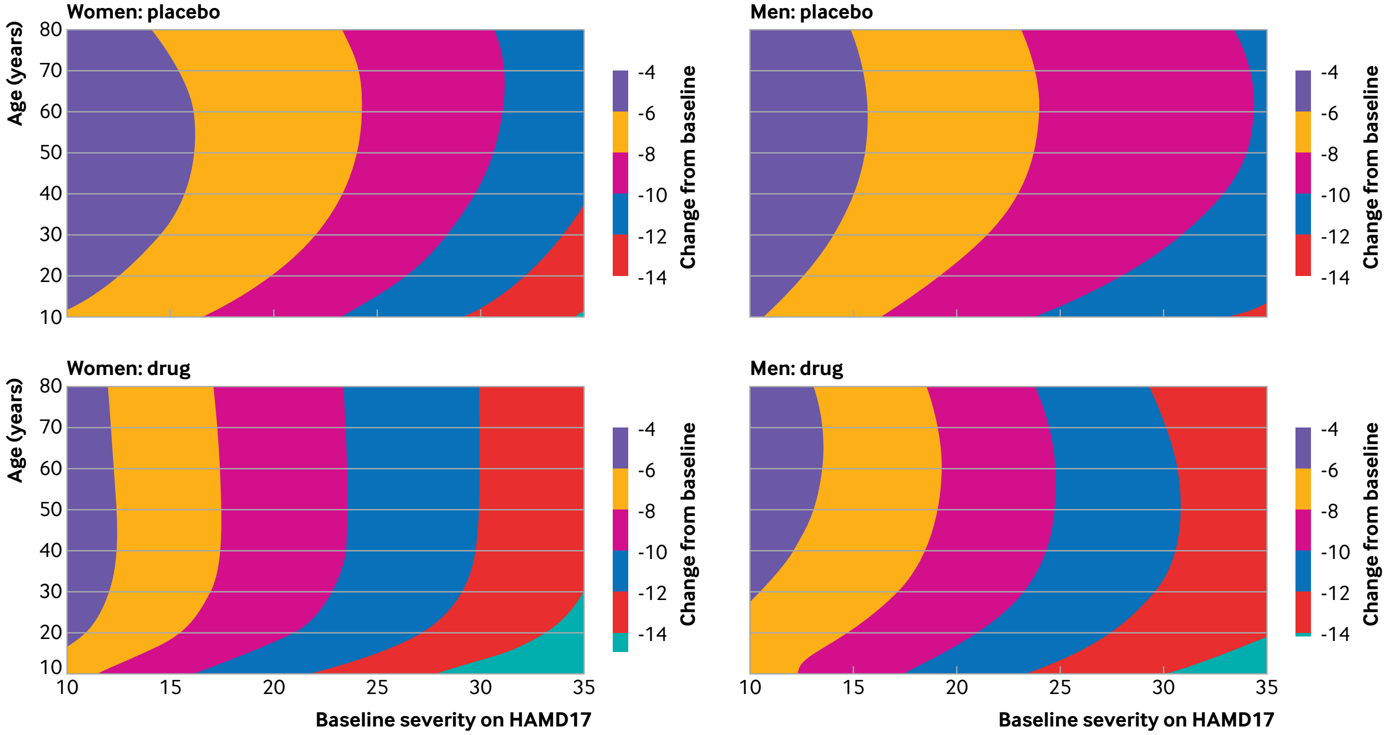
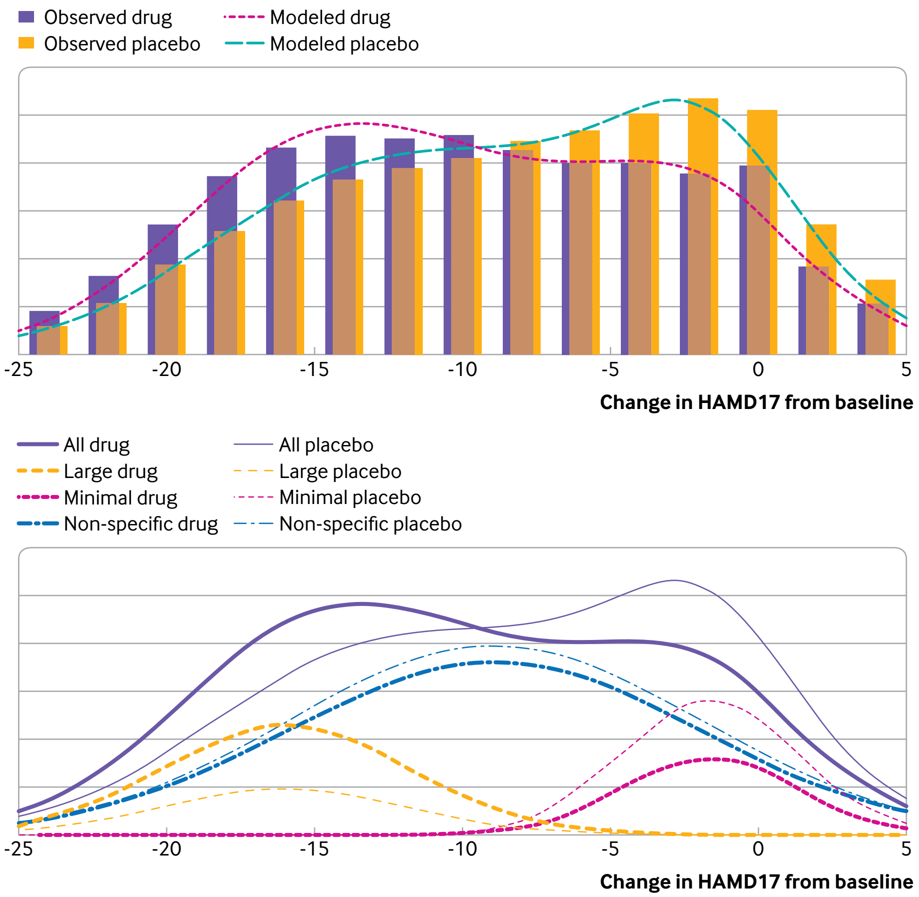


Fig 1 | Heatmaps showing predicted treatment responses (change from baseline) as a function of sex, age, and baseline severity on the Hamilton rating scale for depression (HAMD17)



Distribution	Mean change from baseline	SEM	SD	Estimated proportion of overall population			
				Active drug	95% CI	Placebo	95% CI
Large	-16.00	0.22	4.22	24%	20% to 29%	10%	6% to 13%
Minimal	-1.68	0.11	2.99	12%	10% to 14%	21%	19% to 24%
Non-specific	-8.94	0.15	6.96	63%	58% to 69%	69%	65% to 73%

Fig 2 | (Top) Fit of the mixture model distributions (curves) for drug and placebo responses with the respective histograms for the observed drug and placebo responses. (Bottom) Overall finite mixture model and component normal distributions for drug and placebo. HAMD17=Hamilton rating scale for depression; SEM=standard error of the mean; SD=standard deviation

Strengths and limitations

A key strength of the study was its reliance on a comprehensive dataset free of publication bias,[35] including published and unpublished data. Our findings are limited to acute efficacy while on treatment by the nature of the trials evaluated and the available subject level data; with only observations for baseline and end of treatment, we could not look at symptom trajectories in this study, or other relevant targets of estimation, such as expected symptom severity at a predefined time point (eg, symptom status two months after the start of treatment).

The database did not include details of study design, such as inclusion and exclusion criteria, and their effect on our findings cannot be assessed. Other limitations include lack of more demographic information, evaluation of blinding,[36] individual length of treatment, and item level data that might look at or refine interpretation of modeled subpopulations and the clinical generalizability of the study sample. For example, the effects of treatment history of antidepressants[26] or discontinuations before study entry on our results are unknown, and drug registration trials usually exclude participants with recent suicidal ideation or suicidal attempts, or major medical or psychiatric comorbidities. The patients included in our analysis are thus likely to have had less clinically complex but more acutely severe depression than is typically seen in the community. The STAR*D (Sequenced Treatment Alternatives to Relieve Depression) trial found that about 78% of patients with major depressive disorder are excluded from typical clinical trials. The mean response to antidepressant medication in the first phase of the STAR*D trial was substantially (3.3 points) smaller than that seen in our data, but this finding also does not necessarily indicate that drug effects were smaller.[37]

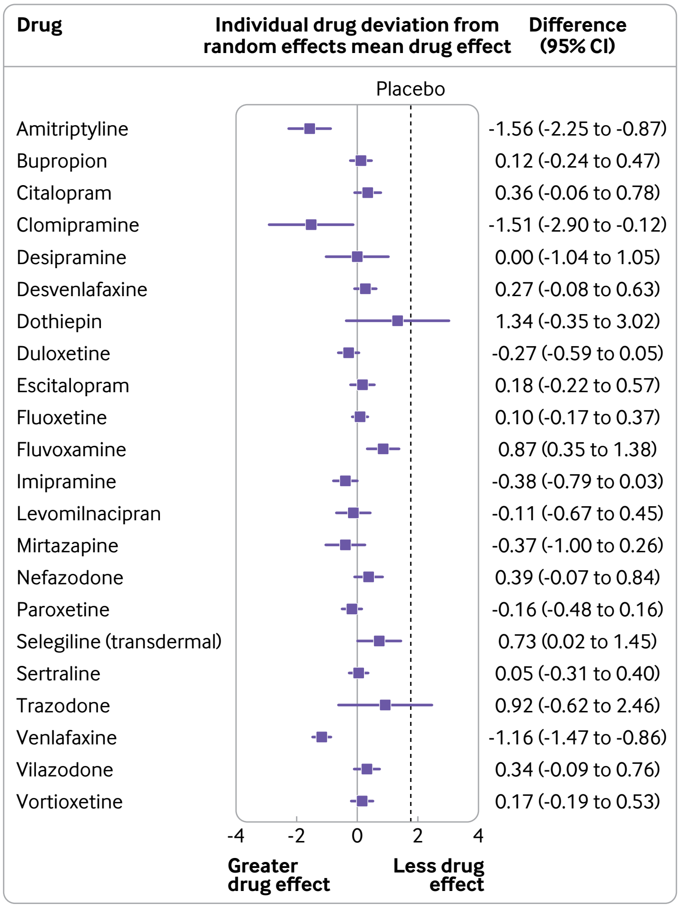


Fig 3 | Estimated effect for each drug. Center line shows the overall (random effects) mean effect for active drug to show how each drug differs from the average. Placebo line is a reference for how the differences among drugs compare with the differences between drugs and placebo

We cannot fully exclude the possibility that the effects of the drugs are accounted for by functional unblinding.[36] This possibility could result in biased assessments by unblinded raters or an increased placebo effect in unblinded participants. If bias were the primary driver of drug effects, we might expect shifts in the means of the response distributions for active drug relative to placebo, or additional response modes, limited to active drug, generated by the subset of participants who were unblinded. Such effects were not seen, however. On the other hand, although the observed distributions cannot exclude drug effects being accounted for by an increased placebo effect because of functional unblinding, drugs with more marked unblinding potential, such as trazodone, mirtazapine, and bupropion, would be expected to evidence larger mean treatment effects than others, but this effect was not seen (fig 3). Finally, HAMD17 has been criticized as a method of assessing changes in depressive symptoms in clinical trials. We found that the Montgomery-Asberg Depression Rating Scale and the Children's Depression Rating Scale correlated strongly with the HAMD17, however, and studies that used the HAMD17 and other scales produced estimates nearly identical in magnitude.

Conclusions

Patients with depression are likely to improve substantially from acute treatment of their depression with drug or placebo. Although the mean effect of antidepressants is only a small improvement over placebo, the effect of active drug seems to increase the probability that any patient will benefit substantially from treatment by about 15%. Further research is needed to identify the subset of patients who are likely to require antidepressants for substantial improvement. The potential for substantial benefit must be weighed against the risks associated with the use of antidepressants, as well as consideration of the risks associated with other treatments that have shown similar benefits.[38-40] Because the benefits and risks might be categorically different (eg, reduced sadness *v* anorgasmia), weighting should be done at the individual level, jointly by patients and their care providers.

This article reflects the views of the authors and should not be construed to represent US Food and Drug Administration's views or policies.

Contributors: MBS (lead author and guarantor) designed and executed the analyses and wrote the manuscript. ZSY contributed to the design of the analyses and wrote the manuscript. BJM contributed to the design of the analyses and the manuscript. KR compiled the data analyzed and contributed to the manuscript. SNK compiled the data analyzed and contributed to the manuscript. IK contributed to the design of the analyses and the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: None.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years except for BJM who reports serving as a member of the CMS Medicare Evidence Development and Coverage Advisory Committee and receiving fees outside the related work from the Federal Trade Commission, the Health Resources and Services Administration, the Heritage Foundation, and Oxidien Pharmaceuticals; BJM was a medical officer at the FDA in the Division of Psychiatry from 2016 to 2017; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required; analysis of aggregated deidentified clinical trial data.

Data sharing: No additional data available.

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: Dissemination of the work to the public and clinical community through social media and lectures is planned.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 James SL, Abate D, Abate KHGBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018;392:1789-858. doi:10.1016/S0140-6736(18)32279-7

2 Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *J Clin Psychiatry* 2015;76:155-62. doi:10.4088/JCP.14m09298

3 Organization for Economic Cooperation and Development. Antidepressant drugs consumption, 2000 and 2015 (or nearest year). Health at a Glance 2017. OECD Indicators, OECD Publishing, Paris. https://doi.org/10.1787/health_glance-2017-graph181-en

4 Pratt LA, Brody DJ, Gu Q. Antidepressant use among persons aged 12 and over: United States, 2011-2014. *NCHS Data Brief* 2017;(283):1-8.

5 Jorm AF, Patten SB, Brugha TS, Mojtabai R. Has increased provision of treatment reduced the prevalence of common mental disorders? Review of the evidence from four countries. *World Psychiatry* 2017;16:90-9. doi:10.1002/wps.20388

6 Weinberger AH, Gbedemah M, Martinez AM, Nash D, Galea S, Goodwin RD. Trends in depression prevalence in the USA from 2005 to 2015: widening disparities in vulnerable groups. *Psychol Med* 2018;48:1308-15. doi:10.1017/S0033291717002781

7 Arroll B, Elley CR, Fishman T. Antidepressants versus placebo for depression in primary care. *Cochrane Database Syst Rev* 2009;(3):CD007954. doi:10.1002/14651858.CD007954

8 Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* 2002;22:40-5. doi:10.1097/00004714-200202000-00007

9 Undurraga J, Baldessarini RJ. Randomized, placebo-controlled trials of antidepressants for acute major depression: thirty-year meta-analytic review. *Neuropsychopharmacology* 2012;37:851-64. doi:10.1038/npp.2011.306

10 Watanabe N, Omori IM, Nakagawa A. Mirtazapine versus other antidepressive agents for depression. *Cochrane Database Syst Rev* 2011;(12):CD006528. doi:10.1002/14651858.CD006528.pub2

11 Cipriani A, Koesters M, Furukawa TA. Duloxetine versus other anti-depressive agents for depression. *Cochrane Database Syst Rev* 2012;10:CD006533. doi:10.1002/14651858.CD006533.pub2

12 Cipriani A, Furukawa TA, Salanti G. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* 2018;391:1357-66. doi:10.1016/S0140-6736(17)32802-7

13 Jakobsen JC, Katakam KK, Schou A. Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and trial sequential analysis. *BMC Psychiatry* 2017;17:58. doi:10.1186/s12888-016-1173-2

14 Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008;5:e45. doi:10.1371/journal.pmed.0050045

15 Gibbons RD, Hur K, Brown CH, Davis JM, Mann JJ. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry* 2012;69:572-9. doi:10.1001/archgenpsychiatry.2011.2044

16 Fournier JC, DeRubeis RJ, Hollon SD. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 2010;303:47-53. doi:10.1001/jama.2009.1943

17 Moncrieff J, Kirsch I. Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. *Contemp Clin Trials* 2015;43:60-2. doi:10.1016/j.cct.2015.05.005

18 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* 2022;27:69-73. doi:10.1136/bmjebm-2020-111600

19 Evers AWM, Colloca L, Blease C. Implications of placebo and nocebo effects for clinical practice: expert consensus. *Psychother Psychosom* 2018;87:204-10. doi:10.1159/000490354

20 Maslej MM, Furukawa TA, Cipriani A. Individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry* 2021;78:490-7. doi:10.1001/jamapsychiatry.2020.4564

21 Volkmann C, Volkmann A, Müller CA. On the treatment effect heterogeneity of antidepressants in major depression: A bayesian meta-analysis and simulation study. *PLoS One* 2020;15:e0241497. doi:10.1371/journal.pone.0241497

22 Plöderl M, Hengartner MP. What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. *BMJ Open* 2019;9:e034816. doi:10.1136/bmjopen-2019-034816

23 Rabinowitz J, Werbeloff N, Mandel FS, Menard F, Marangell L, Kapur S. Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. *Br J Psychiatry* 2016;209:427-8. doi:10.1192/bjp.bp.115.173906

24 Furukawa TA, Maruo K, Noma H. Initial severity of major depression and efficacy of new generation antidepressants: individual participant data meta-analysis. *Acta Psychiatr Scand* 2018;137:450-8. doi:10.1111/acps.12886

25 Hieronymus F, Lisinski A, Nilsson S, Eriksson E. Influence of baseline severity on the effects of SSRIs in depression: an item-based, patient-level post-hoc analysis. *Lancet Psychiatry* 2019;6:745-52. doi:10.1016/S2215-0366(19)30216-0

26 Hunter AM, Cook IA, Tartter M, Sharma SK, Disse GD, Leuchter AF. Antidepressant treatment history and drug-placebo separation in a placebo-controlled trial in major depressive disorder. *Psychopharmacology (Berl)* 2015;232:3833-40. doi:10.1007/s00213-015-4047-2

27 Uher R. Genes, environment, and individual differences in responding to treatment for depression. *Harv Rev Psychiatry* 2011;19:109-24. doi:10.3109/10673229.2011.586551

28 Rutherford BR, Roose SP. A model of placebo response in antidepressant clinical trials. *Am J Psychiatry* 2013;170:723-33. doi:10.1176/appi.ajp.2012.12040474

29 Stahl SM, Greenberg GD. Placebo response rate is ruining drug development in psychiatry: why is this happening and what can we do about it?*Acta Psychiatr Scand* 2019;139:105-7. doi:10.1111/acps.13000

30 Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med* 2017;36:855-75. doi:10.1002/sim.7141

31 Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. *Stata J* 2007;7:45-70. doi:10.1177/1536867X0700700103.

32 National Institute for Clinical Excellence. Depression in Adults (update). 2009. https://www.nice.org.uk/guidance/cg90/documents/depression-in-adults-update-full-guideline-prepublication2

33 Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression: variable, substantial, and growing. *JAMA* 2002;287:1840-7. doi:10.1001/jama.287.14.1840

34 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. doi:10.1192/bjp.bp.111.093336

35 Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252-60. doi:10.1056/NEJMsa065779

36 Baethge C, Assall OP, Baldessarini RJ. Systematic review of blinding assessment in randomized controlled trials in schizophrenia and affective disorders 2000-2010. *Psychother Psychosom* 2013;82:152-60. doi:10.1159/000346144

37 Kirsch I, Huedo-Medina TB, Pigott HE. Do outcomes of clinical trials resemble those "real world" patients? A reanalysis of the STAR*D Antidepressant Data Set. *Psychol Conscious* 2018. doi:10.1037/cns0000164.

38 Khan A, Faucett J, Lichtenberg P, Kirsch I, Brown WA. A systematic review of comparative efficacy of treatments and controls for depression. *PLoS One* 2012;7:e41778. doi:10.1371/journal.pone.0041778

39 Kirsch I, Ness AR, Appleton KM. Treatments for depression: side-effects, adverse events and health risks. *J Affect Disord* 2019;259:38-9. doi:10.1016/j.jad.2019.08.018

40 Gartlehner G, Gaynes BN, Amick HR. Comparative benefits and harms of antidepressant, psychological, complementary, and exercise treatments for major depression: an evidence report for a clinical practice guideline from the American College of Physicians. *Ann Intern Med* 2016;164:331-41. doi:10.7326/M15-1813

Web appendix: Supplementary material