Methods to Reduce Placebo Response in Antidepressant Treatment Trials

PEER REVIEWED

Exaggerated and highly variable placebo response remains one of the principal challenges in the development of drugs for major depressive disorder (MDD), resulting in numerous failed and negative trials. Recent FDA guidance suggests that substantial responses are typically seen in placebo groups in antidepressant trials, and that these responses are often larger than the drug-placebo differences, citing that trials of even effective antidepressants have a failure rate of approximately fifty percent[1]. This high failure rate has resulted in many companies abandoning their depression programmes despite the fact that many patients do not respond adequately to currently available therapies and that depression remains the single leading cause of disability worldwide, and a major contributor to the overall global burden of disease. The World Health Organization estimates that depression affects over 300 million people worldwide or roughly 4.4% of the global population[2]. Various explanations for the rising placebo response rates and inconsistent treatment effect sizes in antidepressant trials have been proposed, including poor study design and analytic methods, various characteristics of the eligible trial population, investigator bias in patient recruitment and selection, the availability of new somatic and non-somatic treatments, and non-specific supportive features of the treatment milieu. Several techniques have shown utility in decreasing exaggerated placebo response in trials of major depressive disorder (MDD) including the following: 1) adopting design options to boost treatment sensitivity and lessen the impact of placebo response; 2) utilising enrichment manoeuvres prior to randomisation; 3) and managing non-specific subject interventions. The following review summarises some of the more salient strategies designed to minimise placebo response in MDD trials.

Design Options

A number of design options can be used to increase the sensitivity to active treatment and potentially lessen the impact of placebo response. This is important as effect sizes (ES) for most antidepressant drugs to date would be considered to be medium according to Cohen's classification, with an average ES of 0.3[3]. One such option to improve ESs is simply to limit the number of treatment arms. Research suggests that reducing the number of treatment arms in a study may positively affect trial outcome, following the observation that most successful trials have fewer treatment arms[4]. This may be because subjects have an increased expectation of receiving active drug in studies where there are relatively more active treatment arms, thereby increasing placebo response. As the likelihood of a subject receiving an effective drug versus placebo has been hypothesised to influence expectation of improvement, and in turn increase placebo response, efficacy studies should try to optimise the number of subjects allocated to placebo. Data suggest that if 50% of subjects are randomised to placebo, the advantage of drug over placebo is expected to be 50% larger than if 25% of subjects are randomised to placebo[5]. For studies with two, three, and four active treatment arms, the maximum probability of at least one significant drug-placebo contrast is achieved with 41.4%, 36.6%, and 33.3% of subjects randomised to placebo, respectively. The basis for this can be appreciated by considering a trial with equal allocation. If one subject is added to the placebo group, the power for all drug-placebo contrasts is increased, whereas if one subject is added to an active treatment arm that is the only contrast where power is increased.

There may also be an advantage to shortening the duration of trials, as placebo response tends to increase with the length of the trial. Khan *et al*. reviewed approval summaries and reported a steadily increasing placebo response over time[3]. The actual length of the trial may not be the significant factor, however. Rather, the number of times the primary outcome measure is administered may have more impact on the placebo response. For example, the general belief is that each time a depression rating scale such as the Hamilton Depression Rating Scale (HAMD) is administered, the total score drops an average of approximately 1.5 points (2 to 4 points on the first two administrations, and 1 to 2 points thereafter). Considering the number of items (potentially hundreds) presented over the course of MDD trials of even relatively short duration, the possibility for desnsitization and the development of a therapeutic alliance and inadvertent therapy becomes apparent. One way to address this is to simply decrease visits or not give the primary outcome measure at each visit. Of note, recent FDA guidance considers six to eight weeks an appropriate study duration for short-term efficacy endpoints for SSRI/SNRI types of antidepressants but for rapid-acting antidepressants, the timing of effect considerations include both efficacy (generally demonstrated within one week for a rapid-acting drugs) and durability over time[1].

In our experience, the use of a single-blind placebo lead-in period has not shown material benefits in reducing placebo response later in the trial, but can be helpful to eliminate patients who are not compliant with taking study drug or completing outcome measures. However, the use of a double-blind, variable placebo lead-in period has shown relatively better sensitivity in detecting and ultimately reducing placebo response in antidepressant trials. Faries *et al*. reported that approximately 28% of patients from a study using a double-blind placebo lead-in met criteria for placebo response, as compared with less than 10% from two single-blind placebo lead-in studies of similar timeframes[6]. More than just confirming earlier notions that the placebo effect is more prominent during the double-blind phase, these data suggest that the important feature in these trials is that the investigator does not know the actual time point of randomization, as this is when investigators change their behavior and subsequent ratings. For this reason cardinal entry criteria such as the minimum cutoff scores for entry into the study should be blinded to both patients and site staff. Randomisation can be accomplished through the use of an algorithm via an automated system and sites merely are told that there will be a minimal level of severity, stability and compliance during the screening/run-in periods and that they will be notified if the patient is eligible for randomisation.

Sequential parallel comparison designs have also shown some merit in reducing placebo response in trials of antidepressants. This is despite some setbacks such as the Euthymics TRIADE MDD trial

in which the active comparator suggested that the study itself had assay sensitivity but that the dose of amitifadine may have been too low. The design was successfully used by Alkermes for its drug ALKS 5461 in a Phase II setting which reportedly showed a very small placebo response of approximately 15%. In the sequential parallel comparison design (SPCD), "the basic idea is to have two phases of treatment. The first phase involves an unbalanced randomisation between placebo and active treatment with more patients randomised to placebo. In the second phase, non-responders treated with placebo are randomised to either active treatment or placebo. Since patients on the second phase have already 'failed placebo,' their placebo response will be reduced" and in this sense this is an enrichment design[7]. The analysis pools the data from both phases in order to maximise power and reduce the required sample size.

Enrichment Manouevres

Several methods, mostly surrounding the assurance of adequately controlled entry criteria can be utilised prior to randomisation in an effort to minimise placebo response. Incorrect classification of diagnosis, which can result from both poorly specified instrumentation as well as clinical biases to enroll subjects who are not appropriate for trials, must be addressed to help decrease rater inflation before randomisation, help minimise placebo response after randomisation, and improve the chances of detecting a drug effect. Over-diagnosis of major depression, which can lead to the inclusion of the temporarily "sad" but otherwise well subjects or subjects depressed secondary to other illnesses, introduces variability in response. This results in an artificial improvement of symptoms post-baseline reflecting regression to the mean. One manouevre to avoid this is to eliminate symptomatic volunteers and subjects whose initial high severity of symptoms can be traced solely to an immediate stressful event.

Another important factor is to ensure that patients have adequate disease severity in order to guard against the over-diagnosis of depression and possible floor effects. Khan has primarily been responsible for a large body of literature suggesting that baseline severity is a prominent success factor in MDD trial sucess[3]. He reported that the overall chance of having a successful trial is only 10% when subjects enter the trial with a mean 17-item HAMD (HAMD-17) <24 at baseline, but there is a 75% chance of success when subjects enter with a mean HAMD-17 >27. Based on subjects' pre-treatment scores on the HAMD-17, he classified subjects into one of four severity groups: low moderate, high moderate, moderately severe, and severe. Associated ESs were 0.51 in the low moderate group, 0.54 in the high moderate group, 0.77 in the moderately severe group, and 1.09 in the severe group. It is possible that subjects with less severe symptomatology may possibly be more prone to the therapeutic milieu and various non-specific factors inherent in MDD trials.

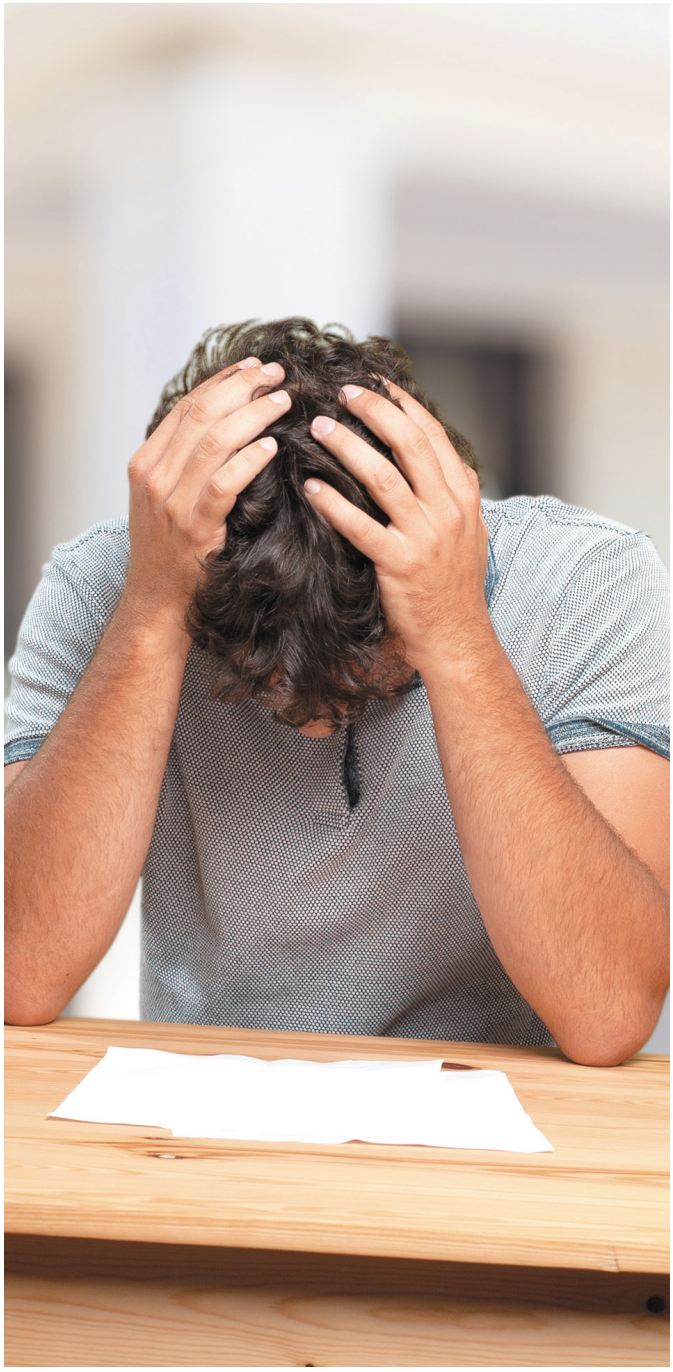
Some reports have suggested that the extent of symptom severity reported by subjects with MDD via an automated version of the Montgomery Asburg Depression Rating Scale (MADRS) and HAMD are equivalent, albeit often lower than those scored by clinicians. There is also some evidence suggesting that the distribution of self-reported baseline scores approximates more of a normal distribution than distributions generated by investigators that tend to be more skewed around the inclusion cutoff scores. This tendency to "fit" subjects to the entry criteria is referred to as rater inflation. Investigators may artificially elevate their ratings in order for patients to meet established severity thresholds and enter the trial inappropriately. In addition to blinding these criteria simply adding threshold scores for several key symptoms not easily distorted or misinterpreted may also help avoid rater inflation. This can include observed symptoms, e.g., emotional dulling and psychomotor slowing, rather than strictly reported symptoms. Another technique would be to employ video-enabled third-party interviews, or other evidence that would be hard to manipulate externally. Other more "objective or solid" evidence of severity such as functional impairment, confirmation of multiple treated episodes, and number of prior hospitalisations can also be utilised to ensure proper patient enrolment.

In addition to severity, stability of symptoms is also an important entry criteria which can be guaranteed by assuming that there would be no more than a 2–3 point change on successive outcome measures, or by only enrolling those patients with relatively lower variance on these measures. Importantly when using a high score on the HAMD or MADRAS (prone to positively biased measurement error) for inclusion purposes, a decrease in score the next time the scale is administered even with no treatment would be expected. This statistical regression to the mean advances that severe or higher scores are more likely to have positive measurement error, while less severe or lower ones are more likely to have negative measurement error. As measurement error by definition is uncorrelated with the true measurement of the underlying construct, the measurement error of any two independent rating scales should be zero. Thus, using one scale for entrance purposes and a different scale for baseline, and then using change scores from this second independent measure should help to statistically eliminate regression to the mean and ultimately help reduce placebo response. In addition, simply temporally separating the screening and evaluative procedures from the administration of treatment may also prove beneficial and help break the therapeutic bond if established.

Non-specific Subject Interventions

The nature of staff-subject interactions, as well as the natural variability in the course of the depression, can produce a clinical setting in which non-specific interactions (e.g., those not attributable to treatment) can appreciably modify a subject's presentation and treatment response. Therefore, it is important for sites to be aware of the therapeutic milieu that they may be engendering. Investigative sites should employ clearly defined and easily monitored conventions that specify the duration, intensity, and type of interactions that might be permitted at all visits. We recommend limiting subject interactions at the site and establishing a codified set of procedures that should be employed by all site staff following Dan Zimbroff's concepts embodied in his Patient and Rater Education of Expectations in Clinical Trials (PREECT) guidance[8]. A key feature of all of these procedures is that every effort should be made to minimise the treatment duration, number of assessments and subject interactions beyond those deemed absolutely mandatory for the successful accomplishment of the study. Although there is insufficient empirical data to suggest that this approach directly results in higher effect sizes, more and more study designs are moving toward an overall simplification of study visits and assessments in MDD trials and antidepressant trials that are designed to answer numerous questions and satisfy disparate audiences in a single setting are decreasing in frequency.

Controlling non-specific subject factors by simplifying trial design, minimising subject interactions and forming a research alliance between site staff and subjects is critical. Site staff and particularly raters should be reminded that they are participating in an experiment and therefore should have no expectations of drug response; either positive or negative. Reminders should also



be made to sites that supportive messages to subjects can adversely affect the sites' ability to provide accurate clinical assessments; that site enthusiasm about the drug may lead to hope and potentially heighten the placebo response; and that this enthusiasm can be especially problematic in start-up activities as these are activities are often designed purposely to build up passion for the drug. Sites should in turn point out to subjects that they are under no obligation to improve during the trial and there should be no stated or unstated communication to subjects that they will improve during the trial. Staff should try to minimise conversation that is overly empathic, placating or encouraging, and communicate with subjects in a neutral manner while ensuring that all interactions across subjects are as consistent as possible. This does not mean that site staff should behave as automatons; rather a middle ground that is neither overly supportive nor overtly emotionless should be sought. The ethics of clinical research requires true clinical equipoise on the part of the site staff, which is a state of genuine uncertainty regarding the treatment effectiveness of the trial. A program designed to assess site staff's beliefs and expectations about research and then help reduce variability and ultimately placebo response should be a basic part of all MDD clinical trials.

.

REFERENCES

1. Draft Guidance for Industry: Major Depressive Disorder Developing Drugs for Treatment. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER). June 2018. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM611259.pdf
2. World Health Organization. Depression and Other Common Mental Health Disorders Global Health Estimates. 2017. WHO Ref # WHO/MSD/MER/2017.2. http://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/
3. Khan, A. and Brown, W. Antidepressants versus placebo in major depression: an overview. World Psychiatry 2015;14:294-300)
4. Khan, A., Broadhead, A., Kolts, R. and Brown, W. Severity of depressive symptoms and response to antidepressant and placebo in antidepressant trials. J Psychiatric Research, 2005, 39 (2); 145-50.
5. Mallinckrodt, C., Tamura R. and Tanaka, Y. Recent developments in improving signal detection and reducing placebo response in psychiatric clinical trials. Journal of Psychiatric Research 45 (2011) 1202-1207.
6. Faries, D., Heiligenstein, J., Tollefson, G. and Potter, W. The double-blind variable placebo lead-in period: results from two antidepressant clinical trials. J Clin Psychopharmacol. 2001 Dec; 21(6): 561-568.
7. Fava, M., Evins, E., Dorer, J. and Schoenfeld, D. The Problem of the Placebo Response in Clinical Trials for Psychiatric Disorders: Culprits, Possible Remedies, and a Novel Study Design Approach. Psychother Psychosom 2003; 72: 115–127.
8. Zimbroff DL. Patient and rater education of expectations in clinical trials (PREECT). J Clin Psychopharmacol. 2001 Apr; 21(2): 251-2.

Henry J. Riordan

Henry J. Riordan, Ph.D. is Executive Vice President of Scientific Solutions at Worldwide Clinical Trials. Dr Riordan has been involved in the assessment, treatment and investigation of various neuroscience drugs and disorders in both industry and academia for the past 20 years. He has advanced training in neuroimaging, neuropsychology, experimental design and statistical methodology. He has over 100 publications, including co-authoring two books focusing on innovative CNS clinical trials methodology.

Email: henry.riordan@worldwide.com

Rolana Avrumson

Rolana is the Director of the Clinical Assessment Technology (CAT) group at Worldwide. Her responsibilities include rater training services, subject eligibility reviews, data surveillance, scale management and spearheading programs to reduce placebo response. As the leader of the rater training team she has extensive experience in administering, scoring, and analyzing psychiatric and cognitive measures across a number of pediatric, adult, and geriatric CNS populations spanning various neurologic, psychiatric and analgesic indications.

Email: rolana.avrumsom@worldwide.com