When Placebo Response Isn’t Placebo
Response: THE PERILS OF HUMAN RATERS
by Dr. Jack Modell, Senior Medical Officer, Rho

PRIOR TO JOINING the pharmaceutical and contract research organization industries, I was in clinical practice for twenty years as a psychiatrist and medical researcher. And something I noticed very early on among my patients with major mental illnesses, particularly those with severe depression and psychotic disorders, was that they did not generally get better — at least not for more than a day or two — by my simply being nice to them, treating them with ineffective medications (e.g., vitamins when no vitamin deficiency existed), seeing them weekly for office visits, or by providing other so-called supportive interventions that did not directly address the underlying illness. To be clear, this is not to say that kindness and supportive therapy are not critical to the patient-physician relationship (“The secret of the care of the patient is in caring for the patient” [Frances Weld Peabody, 1927]), but rather that kindness and support alone rarely make a biologically based illness substantially improve or disappear.

With this background, I vividly recall my surprise upon being asked shortly after I joined the pharmaceutical industry: “Can you help us figure out how to decrease the nearly 50% placebo-response rate we see in antidepressant trials for major depressive disorder?” “Fifty percent?” I replied, incredulously. “There’s no way that 50% of patients in a true antidepressant trial for major depressive disorder would show me voluminous data supporting their figure.

I spent the next few years trying to figure out this apparent paradox. Not surprisingly, the answer turned out to be manifold. After careful review of internal and external data, as well as published explanations for high “placebo response rates” in clinical depression trials (much of which also applies to clinical trials in general), the following three factors emerged as being of particular importance because they are easily mitigated by proper trial design and research staff training.

1) Subjects being admitted into clinical trials often had depressive symptoms, but did not truly meet criteria for major depressive disorder. Examples include subjects with personality disorders whose symptoms wax and wane considerably with external factors (e.g., family or job stress), subjects with depressive symptoms in response to a particular stressor (not of sufficient severity or duration to meet formal criteria for a major depressive episode and likely to abate with the passage of time), and subjects who — for various reasons — may feign or exaggerate symptoms for the purpose of seeking attention or gaining access to a clinical trial. Unlike the patients I encountered in my clinical practice, subjects with these presentations often do improve with supportive interventions and placebo.

Recruitment of truly depressed subjects is made even more difficult by the widespread availability of reasonably effective medication options. Patients in the throes of a major depressive disorder, who sometimes have difficulty even making it through the day, are rarely keen to commit to the additional efforts, uncertainties, and treatment delays involved with a clinical trial when an inexpensive prescription for an effective generic antidepressant can now be filled in a matter of minutes. Indeed, as more and more generally safe and effective medications have become approved and readily available for a variety of illnesses, the motivation for patients to join clinical trials in the hope of finding an effective treatment has correspondingly decreased.

2) The second factor is somewhat difficult to discuss because it sometimes provokes an understandable defensive response in clinical investigators. Consciously or unconsciously, many investigators (which includes clinical raters) inflate or deflate clinical ratings to enable the subject to gain entry into, or remain enrolled in, a clinical trial. Most commonly, this is done by subtly (or not so subtly) coaching subjects on their answers or, when subject responses or findings seem to fall in between scale severity ratings, by rounding up or down to a rating that is more likely to qualify the subject for the trial.

The effect of this practice is diagrammed in the following figures, specific examples of which can be seen in these references.1,2 In figure 1, the white bell-shaped distribution is the expected distribution in severity rating scores of an unselected clinical population presenting for clinical trial participation, let’s say with a mean score at shown at X. Not uncommonly, what we see in clinical trials in which a certain scale severity score is required for study entry (depicted by the vertical light blue line) is not the expected right half of this bell-shaped distribution, but rather a distribution like that shown by the orange curve, which is essentially the right-half of the bell-shaped distribution with a large proportion of subjects whose ratings fell to the left of the blue line (i.e., those falling short of required severity for study entry) “pushed” to the right, over the blue line, so that they now qualify for study inclusion, with the mean of those thus selected shown at X.

At the first follow-up visit, when raters (and subjects) now have little incentive to influence rating scores to meet a pre-specified criterion, the scores of the entire included population are free to relax towards their true values and assume the pre-selection and more normally distributed pattern. Moreover, subjects and investigators, expecting that the onset of treatment should coincide with at least some clinical improvement, may bias rating scores during this period to reflect this expectation even though the signs and symptoms of the illness may have yet to show true change. During this same time, any actual clinical improvement will also result in the rating score mean shifting leftward (white arrow, figure 2), but because the measured change — from the initial X of the selected population to the

---

Figure 1

---

Figure 2
new mean (x̄n1; orange arrow, figure 2) – is generally much greater than a true treatment effect during this period, any real changes are obscured and the ability to detect a true drug-placebo difference may be lost. While this early “improvement” in rating scores for subjects in clinical trials may appear to be a “placebo effect” and is often confused with it, this apparent improvement is instead the result of artificially inflated scale scores regressing back to their original true distribution, in combination with whatever actual treatment and placebo effects may have occurred during this time. Unfortunately, the introduction of non-qualified subjects to the study and rater bias will continue to hamper detection of actual drug-placebo differences throughout the course of the study.

3) Finally, investigators and site staff often do not fully understand the true objective of the clinical trial: it should never, for example, be “to show treatment efficacy”, but rather, to test the null hypothesis of no treatment difference or to estimate likely treatment effect, as well as to faithfully and objectively record all adverse effects that may emerge during treatment. Likewise, they often fail to understand the importance of complete objectivity and consistency in performing clinical ratings, the intention behind and importance of every inclusion and exclusion criterion (necessary for their proper interpretation and application), and the destructive effect on the outcome and scientific integrity of the trial that even well-intended efforts to include subjects who are not fully qualified can have.

Each of these three factors can skew both drug and placebo trial populations and results, making it appear that subjects “improved” well beyond what would have resulted with strict adherence to protocol requirements and objective assessment of study entry and outcome measures.

What, then, can be done to prevent these problems from sabotaging the results of a clinical trial? Foremost are thorough and meticulous investigator and rater education and training. All too often, perfunctory explanations of the protocol and clinical assessment tools are provided at investigator meetings, and “rater training” takes the form of brief demonstrations of how the rating scales are used and scored, without actually testing raters to be certain that they fully understand how the scales are to be used and interpreted, including understanding scoring conventions, criteria, and necessary decision-making. Even seemingly sound training has marked limitations both immediately and as training effects deteriorate during conduct of the trial.

Training of the research staff must include not only a review of the protocol design and study requirements, but detailed explanations about why the trial is designed exactly as it is, the importance of strict adherence to study inclusion and exclusion criteria, and the necessity for complete honesty, objectivity, and consistency in conducting the clinical trial and in performing clinical assessments. Detailed training on the disease under study is also important to ensure that all site staff have as complete an understanding as possible of the intended clinical population and disease being studied so that they can assess subjects accordingly. And, as noted above, rater training must include not only education on the background, purpose, characteristics, and instructions for each scale or outcome measure used, but trainers, as well as investigators and raters, should be tested for adequate understanding and proficiency in use of each of these measures.

Finally, and perhaps most importantly, is that no matter how well-informed or well-intentioned investigators and raters might be, humans simply cannot match computers in objectivity and consistency, including of the kind needed to make assessments based on subject responses to questions in clinical trials. Unless being programmed to do so, a computer cannot, for example, coach a subject on how to respond, nor would it inflate (or deflate) ratings based on feelings, expectations, response interpretations, or desired outcomes. A computer faithfully asks the same questions every time, following the same algorithm, and records responses exactly as provided by the subject. Indeed, several studies have shown that computerized assessments of entry criteria and outcome measures in clinical trials – in particular interactive voice response systems (IVRS) and interactive web response systems (IWRs) – provide data of quality and signal-detection ability that meet and often exceed that obtained by human raters. For these reasons, strong consideration should also be given to using IVR and/or IWR systems for assessing study entry criteria and endpoints that allow such use.

The author gratefully acknowledges John H. Greist, MD, for his outstanding research and input regarding these important findings and considerations.

References

Dr. Jack Modell, Rho’s Senior Medical Officer, is a board-certified psychiatrist with 20 years of experience in clinical research, teaching, and patient care in academic medicine. He has over 10 years of experience in clinical drug development (proof of concept through market support), medical affairs, successful NDA filings, medical governance, drug safety, compliance, and management within the pharmaceutical and CRO industries. Jack’s specialties and expertise include neuroscience, clinical pharmacology, drug development, clinical research, medical governance, and clinical diagnosis and treatment.