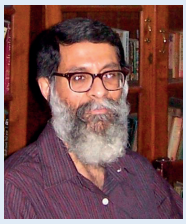
Signal-to-Noise Ratio, Variability, and Their Relevance in Clinical Trials

Chittaranjan Andrade, MD



Each month in his online column, Dr Andrade offers practical knowledge, ideas, and tips in psychopharmacology to *JCP* readers in psychiatric and general medical settings.

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bangalore, India (candrade@psychiatrist.com).

Dear Reader,
Don't miss out on the many peer-reviewed offerings that we publish only online.
Visit *JCP* online to read more by Dr Andrade:

- Breast Cancer and Antidepressant Use
- Modafinil and Armodafinil in Schizophrenia
- Drugs That Escape Hepatic Metabolism
- Schizophrenia and Smoking
- SSRIs and Persistent Pulmonary Hypertension of the Newborn

Visit
PSYCHIATRIST.COM
Enter Keyword PRACTICAL

J Clin Psychiatry 2013;74(5):479–481
(doi:10.4088/JCP.13f08475)

Clinical Questions

Clinicians are often puzzled by questions such as the following:

1. Why do large, double-blind, randomized controlled trials (RCTs) so often fail to yield a statistically significant result favoring the trial medication? This question appears particularly applicable to antidepressant trials in both adult and pediatric populations, even with drugs that are later approved by regulatory authorities.[1,2]
2. Why are there so many small RCTs with statistically significant results? One would expect that with smaller samples it would be more difficult, not easier, to achieve statistical significance.
3. Why is the number needed to treat (NNT) figure so large in antidepressant and other RCTs[3,4]? Do medications really play such a small role in eliciting treatment response?

These questions have several answers. For example, RCTs often fail because clinical trials are designed and executed in a way that inevitably recruits a high placebo response.[5] Because there is a limit to how much patients receiving an experimental drug can expect to improve, enhancement of response in the control group will shrink the difference in outcomes between drug and placebo, predisposing to statistically nonsignificant outcomes. Or, if the difference between groups is indeed statistically significant, the NNT is large because it took a really large sample in a multicenter study to show a small advantage for the experimental drug. Finally, small RCTs seemingly more often have significant outcomes because there may be an equal number of small RCTs with negative outcomes that have not been published, perhaps because authors or journal editors were insufficiently interested in the results. These and other explanations have been discussed elsewhere.[6,7] A clear, practical discussion on the subject is also available on the Web.[8]

With regard to the clinical questions listed above, the present article will introduce a concept that has received little attention in psychiatric literature. This concept relates to noise in the measurement of outcome variables and the signal-to-noise ratio in research. This article will also discuss how variability in the signal affects RCT outcomes.

The Signal-to-Noise Ratio

The signal-to-noise ratio is an important concept in several different sciences, including medicine. The concept is quite simple. Imagine that you are at a party where many groups of people are talking and laughing over the sound of raucous rock music. If you speak to your friends in a normal conversational voice, whether or not you are heard well will depend on the signal-to-noise ratio, that is, how well your "signal" can be clearly discerned amid the general background "noise."

Event-related potentials (ERPs) provide an excellent illustration of the signal-to-noise ratio in neuropsychiatry. In ERPs, an event such as a sensory stimulus triggers a small spike in the electroencephalogram (EEG) in a part of the brain. This spike cannot be easily identified because it is submerged in the background brain electrical activity. The spike is the signal, and the background EEG activity is the noise. Now, if the sensory stimulus is repeated hundreds of times and if the EEG activity is recorded and averaged, the signal becomes easier to identify because it is a consistent response, whereas

- Response to medication is an example of a "signal" that clinicians and researchers wish to detect and measure.
- Medication response varies across patients. Some of the variability is genuine and is due to reasons that are intrinsic to the patient. Some of the variability is spurious and is due to measurement error or "noise."
- This article explains the concept of the signal-to-noise ratio as applied to clinical research and practice. Examples of sources of noise are provided. Suggestions are made for how to reduce noise in clinical assessments.

the background EEG activity, being relatively and contextually random, tends to cancel out.[9]

The Signal-to-Noise Ratio in Clinical Research and Practice

In clinical trial research, the signal is anything that the investigator is trying to measure. Noise is anything that produces a smear in the value of the signal, making it hard to identify the true value of the signal. For example, in a clinical trial of a new antidepressant, severity of depression is a signal of interest. This signal needs to be accurately measured as distinct from the background noise. Examples of sources of background noise that could interfere with the accurate measurement of this signal are a quarrel that the patient had that affects his mood, the long journey to the hospital that increases his fatigue, the anticipated discomfort associated with blood sampling that bumps up his anxiety levels, the reassurance provided by the clinical team that makes him feel better, and the nonprescription medication he is taking that improves his sleep. Readers may note that noise of this nature may interfere with interpretations of clinical status in everyday clinical practice as well, much as it does in clinical trials.

The real signal, which explains how the antidepressant has affected his depression ratings, must be picked out from all the background factors that add noise to the signal. Thus, in clinical research as in ERP studies, the background noise merges with the signal and requires averaging across a large number of participants for the signal to be identified. The greater the noise, the smaller the signal-to-noise ratio and the larger the sample size necessary to identify the signal.

Noise, Variability, and Their Effects in RCTs

Noise in clinical trials increases the variability in the value of the signal. Different patients will experience different sources and intensities of noise; therefore, the magnitude and direction of the noise-induced variation will vary across patients, as should be obvious from the examples provided earlier. The greater the variation, the more difficult it becomes to identify a signal and hence the more difficult it is to demonstrate statistical significance. This phenomenon can be understood from the hypothetical and considerably exaggerated data presented in Table 1. The numbers in the table are the depression scores in different patients at the end of 2 RCTs that compared the same antidepressant drug with placebo. Patients in group 1 received drug, and those in group 2 received placebo. It is intuitively apparent that the patients in group 1 improved more than did those in group 2; that is, drug was superior to placebo.

Table 1. Depression Scores for Individual Patients at the End of 2 Hypothetical RCTs[a]

						Mean Score
RCT 1						
Group 1 (antidepressant)	14	14	15	16	16	15
Group 2 (placebo)	21	21	22	23	23	22
RCT 2						
Group 3 (antidepressant)	5	10	15	20	25	15
Group 4 (placebo)	2	12	22	32	42	22

[a]Both RCTs studied the same antidepressant. There were 5 patients in each group in each RCT. Each value represents the endpoint depression score of a single patient. The group means and difference between group means are the same in the 2 RCTs. Whereas the homogeneity in scores within groups in RCT 1 suggests that the groups differ significantly, the large variation and considerable overlap of scores between groups in RCT 2 imply that the groups do not differ significantly.
Abbreviation: RCT = randomized controlled trial.

Unfortunately, there was greater noise in the second RCT, and, although the mean values for and the difference between groups 3 (drug) and 4 (placebo) remain the same, the superiority of drug over placebo is no longer convincing, because there is greater variability within groups and hence substantial overlap of scores between groups.

From a statistical perspective, greater variability due to noise means that, for a given sample size, it becomes harder to detect the signal and hence a statistically significant outcome. In such situations, the sample size needs to be larger to detect the signal that the antidepressant drug results in lower depression ratings than placebo, if indeed this is the case.

True Variability and Noise

Noise is anything that introduces unreliability in the rating of a variable in a study. Noise increases the variability in the signal, but not all variability in the signal is due to noise. Expressed otherwise, variability can be true or spurious; the latter reflects an inaccurate estimation of the signal and is due to noise.

True variability arises from sources that are intrinsic to the patient, such as genetically driven or age-driven differences in antidepressant response, differences in the ways in which men and women experience depression, differences in the nature and degree of the stressors that maintain the depression, and so on. The signal is truly different across individuals. In a multicenter RCT, true variability can also arise from genuine population differences across sites.

Spurious variability arises from sources that are external to the patient, such as differences in the ways in which patients are recruited, managed, and rated. In single-site studies, the same investigators usually run the study from beginning to end, and the same standard operating procedures are applied in a similar fashion to all recruited patients. So, external biases (if any) in recruitment, management, and rating tend to move all ratings of the signal in the same way, and the variability may not be much increased. If variability is not amplified, smaller samples may suffice to identify statistically significant outcomes, if any.

In multicenter studies, despite attempts to standardize operating procedures across sites, there could be intersite differences in the thresholds for recruitment, differences in the thresholds for advising rescue medications, differences in nonspecific psychotherapeutic support, differences in the ways in which patients are rated, and even differences in decision-making related to

Table 2. A Few Suggestions for Reduction of Spurious Variability (noise) in Clinical Ratings

1. Ensure that structured, standardized instruments are employed.
2. Ensure that the raters are experienced in assessing patients with the psychopathology of interest, so that they can efficiently recognize the signal.
3. Ensure that the raters have considerable previous familiarity with the rating instrument so that they can use it competently to obtain an accurate measurement of the signal.
4. Ensure that if there is more than 1 rater, all raters are assigning scores in the same way; this rater training may need to be repeated across the course of a long study to ensure that rater drift does not occur.
5. Ensure that rating is conducted at the same time of day so that diurnal variations in illness do not prejudice the ratings, and ensure that rating is conducted in the same sequence with regard to other study procedures so that prior biasing influences, if any, are uniform.

removing patients from the study (resulting in differences in last-observation-carried-forward values). As a result, there is a potential for a considerable increase in the variability of the outcomes that are rated. For all these reasons, the signal is vulnerable to being smeared in multicenter RCTs. The resultant increase in variability and fall in the signal-to-noise ratio make it harder for the study to identify statistically significant outcomes. Thus, a large sample is necessary to pick out the signal from the noise, and the eventual NNTs are large.

Reducing Noise and Variability: Advantages and Disadvantages

A great deal could be written on how to reduce variability that arises from noise, all of which has to do with uniform observance of standard operating procedures within and across recruitment sites, as well as the maintenance of high standards of rater reliability. A few examples of how to reduce spurious variability are listed in Table 2; for the rest, the reader is referred to standard texts on research methodology and on how to conduct clinical trials.

The suggestions listed in Table 2 all address the reliability of clinical ratings. Does reduction in noise through improvement in reliability really make a different to the conduct of clinical trials? Yes, indeed. For example, over a decade ago, Perkins et al[10] showed that improvement in reliability could meaningfully reduce the sample size necessary for an adequately powered study.

There are also ways in which true variability can be reduced, and all involve recruiting as homogeneous a sample as possible, such as patients of a particular gender, in a particular age group, with a specific diagnosis and subdiagnosis, with a specified severity of illness, with similar history of previous medication exposure, and so on. Controlling environment (inpatient, as opposed to outpatient setting) could also help.

Reduction in noise is definitely desirable because it improves the internal validity of a study and its conclusions. Reduction in true variability, however, will improve the ability of the study to detect the signal but will compromise the external validity of the study because the results may not be easily generalizable to other populations. This is a well-known limitation of regulatory RCTs that exclude patients with personality disorders, substance abuse, comorbidities, and other descriptors that commonly characterize patients seen in everyday clinical practice. Researchers must therefore decide what they want from the study: to detect the signal or to generalize the results of the study.

Animal studies are a particular example of how reduction in variability can be advantageous as well as disadvantageous. Such studies commonly employ (for example) rats that are of the same age, body weight, gender, and inbred strain. The animals are housed and studied in a carefully controlled environment that is free from environmental disturbances. The animals are handled in the same way and are studied at the same time of day. True variability and variability due to noise are simultaneously reduced to a minimum. As a result, it is very easy to pick up even a small signal. However, this same signal could be completely drowned out in clinical trials because of the complex genetic and environmental variability that characterizes patients and because of the noise that is inevitable in clinical trial research. This is one of the many reasons why a finding that is statistically significant in the laboratory may not be statistically significant in clinical contexts.[11]

Parting Note

In the RCT situations discussed, the signal-to-noise ratio was explained in the context of dependent variables, such as the severity of depression in an antidepressant trial. However, the concept applies to the measurement of any variable, whether dependent or independent. For example, noise can contaminate the measurement of blood pressure in a study of how hypertension impacts the risk of Alzheimer's disease in later life. In this example, blood pressure is an independent variable.

References

1. Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med*. 2008;5(2):e45.
2. Andrade C, Bhakta SG, Singh NM. Controversy revisited: selective serotonin reuptake inhibitors in paediatric depression. *World J Biol Psychiatry*. 2006; 7(4):251–260.
3. Arroll B, Elley CR, Fishman T, et al. Antidepressants versus placebo for depression in primary care. *Cochrane Database Syst Rev*. 2009;(3):CD007954.
4. Tsapakis EM, Soldani F, Tondo L, et al. Efficacy of antidepressants in juvenile depression: meta-analysis. *Br J Psychiatry*. 2008;193(1):10–17.
5. Andrade C. There's more to placebo-related improvement than the placebo effect alone. *J Clin Psychiatry*. 2012;73(10):1322–1325.
6. Kobak KA, Kane JM, Thase ME, et al. Why do clinical trials fail? the problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol*. 2007;27(1):1–5.
7. Greist J, Mundt J, Jefferson J, et al. Comments on "Why do clinical trials fail? the problem of measurement error in clinical trials: time to test new paradigms?" *J Clin Psychopharmacol*. 2007;27(5):535–537.
8. Chin R. Why clinical trials fail. Clinicaltrialist Web site. http://clinicaltrialist.wordpress.com/clinical/why-clinical-trials-fail/. Accessed March 18, 2013.
9. Woodman GF. A brief introduction to the use of event-related potentials in studies of perception and attention. *Atten Percept Psychophys*. 2010;72(8): 2031–2046.
10. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry*. 2000;47(8):762–766.
11. Andrade C, Sudha S, Venkataraman BV. Herbal treatments for ECS-induced memory deficits: a review of research and a discussion on animal models. *J ECT*. 2000;16(2):144–156.