Clinical and Practical **Psychopharmacology**

The Primary Outcome Measure and Its Importance in Clinical Trials

Chittaranjan Andrade, MD

Each month in his online column, Dr Andrade considers theoretical and practical ideas in clinical psychopharmacology with a view to update the knowledge and skills of medical practitioners who treat patients with psychiatric conditions.

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

ABSTRACT

The primary outcome measure is the outcome that an investigator considers to be the most important among the many outcomes that are to be examined in the study. The primary outcome needs to be defined at the time the study is designed. There are 2 reasons for this: it reduces the risk of false-positive errors resulting from the statistical testing of many outcomes, and it reduces the risk of a false-negative error by providing the basis for the estimation of the sample size necessary for an adequately powered study. This article discusses the setting of the primary outcome measure, the need for it, the increased risk of false-positive and false-negative errors in secondary outcome results, how to regard articles that do not state the primary outcome, how to interpret results when secondary outcomes are statistically significant but not the primary outcome, and limitations of the concept of a primary outcome measure in clinical trial research.

J Clin Psychiatry 2015;76(10):e1320–e1323
dx.doi.org/10.4088/JCP.15f10377

Introduction

Research articles that describe randomized controlled trials (RCTs) usually but not always specify a primary outcome measure. This article explains what a primary outcome measure is, why it is necessary to specify the primary outcome a priori, how one may interpret the primary and secondary outcomes reported in a research article, and limitations of the concept of primary and secondary outcomes.

Primary and Secondary Outcome Measures

The primary outcome measure is the variable that an investigator considers to be the most important among the many dependent variables that are to be examined in the study. As an example, an investigator plans to compare a new antidepressant drug with placebo in an 8-week RCT in patients with major depressive disorder (MDD). He decides that he will administer the Montgomery-Asberg Depression Rating Scale (MADRS), the Hamilton Depression Rating Scale (HDRS), the Clinical Global Impression (CGI) scales for Severity (CGI-S) and Improvement (CGI-I), and instruments that measure quality of life, sexual functioning, and medication adverse effects. He also plans to record vital physiologic parameters such as the heart rate and blood pressure and obtain electrocardiograms and routine laboratory tests.

All of these are outcome measures, or dependent variables. Out of this long list, the investigator decides that improvement on the MADRS is the most important; if the antidepressant drug attenuates MADRS ratings significantly more than does placebo, he will conclude that the drug is effective in treating MDD. In other words, he sets improvement on the MADRS as the primary outcome in the RCT; the result on this single outcome is the primary determinant of whether the study is considered a "success" or a "failure." All of the remaining assessments are hierarchically less important and comprise the secondary outcomes.[1]

In this RCT, the investigator may choose, instead, to designate response rate (defined as 50% attenuation of MADRS scores) as the primary outcome measure. However, this would not be a good idea because a much larger sample size would be necessary to identify statistically significant differences for categorical outcomes (eg, response rate) as compared with continuous outcomes (eg, reduction in MADRS scores).

In another example, an investigator plans to conduct an RCT that compares risperidone with haloperidol in patients with schizophrenia. He wishes to find out whether risperidone is associated with better cognitive outcomes. He identifies a large battery of neuropsychological tests, including tests of attention, concentration, working memory, logical memory, visuospatial memory, ideational fluency, perceptuomotor speed, and problem solving. He sets a composite cognitive index, formed from mathematically combining the results of these neuropsychological assessments, as the primary outcome measure. Alternatively, to save himself the bother, he could instead designate improvement in working memory as the primary outcome measure.

Once the primary outcome measure has been set, the remaining efficacy and adverse effect outcomes comprise the secondary outcome measures.

Clinical Points

- The primary outcome measure is the outcome that an investigator considers to be the most important among the many outcomes that are to be examined in the study.
- The primary outcome is defined at the time the study is designed. This provides a basis for the estimation of sample size and reduces the risk of false-positive errors resulting from the statistical testing of many outcomes.
- Statistical testing of secondary outcomes is associated with an increased risk of both false-positive and false-negative errors.

In this article, the terms *primary outcome* and *primary outcome measure* are used interchangeably, for convenience. Readers must note, however, that the primary outcome refers to the identified dependent variable, whereas the primary outcome measure as a complete concept additionally includes the time at which the dependent variable is measured (eg, study endpoint), the method of analysis (with covariates, if any, specified, a priori), and the sample on which the analysis will be conducted (eg, the intent-to-treat sample).

Setting the Primary Outcome Measure

The primary outcome measure is set at the time the study is designed and the study protocol is drafted, that is, before the study is begun.[2] There are 2 reasons for this:

1. Setting the primary outcome measure a priori prevents the investigators from cherry-picking significant results and presenting these as the main findings of the study.
2. The primary outcome measure forms the basis for the calculation of the sample size required for the study.

Each of these is discussed in greater detail in subsequent sections.

The primary outcome measure is set by the investigator and the research team based on consensus opinion or based on what previous investigators did in similar studies.[3] Sometimes, study sponsors, or even regulatory bodies, may decide what the primary outcome measure ought to be. The primary outcome measure is the outcome that best decides whether a study is a success or a failure.

Research articles usually define the primary outcome measure in the methods section and sometimes also in the abstract. If the primary outcome is not stated, or if the reader wonders whether the authors have correctly represented the primary outcome in their article, clarification can be obtained from a visit to the clinical trial registry in which the study was registered. This is because the primary outcome that is set is required to be explicitly defined in the study protocol that would have been uploaded into the registry database. It has been known for authors to misrepresent in their articles the primary outcome to which they had committed themselves in their protocols.[4]

The Primary Outcome Measure and Protection Against a Type I Error

A primary outcome measure needs to be defined a priori to protect against the risk of a false-positive error arising from the application of statistical tests to a large number of trial outcomes.[2] Here is the explanation. Setting α for statistical significance at $P < .05$ means that if a conclusion (eg, the treatment is effective) regarding an outcome is not true in the population (that is, the null hypothesis is true), then, if we perform the study a hundred times, it will correctly be found to be untrue approximately 95 times or more and mistakenly (by chance) be found to be true approximately 5 times or less. So, the type I or false-positive error rate is 5% or lower.

Extending the concept, if we have a large number of outcome measures in an RCT, we will be performing a large number of statistical tests. If the null hypothesis is true and we perform 100 tests with α set at 5%, about 5 of these outcomes could be expected to emerge significant by chance. Furthermore, it can mathematically be shown that when α is set at 5%, if k statistical tests are performed, then the risk of at least 1 false-positive result is $(1 - 0.95^k)$. So, if 10 tests are performed with α set at 5%, there is a 40% chance that at least 1 of these will be statistically significant by chance when the conclusion is not true in the population.

In other words, if an investigator does not set a primary outcome measure in advance, he can cherry-pick whatever emerges that is statistically significant as the finding to emphasize in his article, even though the significance might have been a chance outcome. This would be cheating. Requiring a primary outcome measure to be set a priori prevents such cherry-picking.

P is truly set at $< .05$ only for the primary outcome measure. For the secondary outcome measures, given that there will be many of these, the effective false-positive error rate would probably be much higher than 5% even though we think that it is 5%. The larger the number of secondary outcomes, the higher the likely false-positive error rate. This is the reason why, ideally, investigators and readers should pay most attention to the primary outcome measure in the study results and why secondary outcomes should be viewed with caution.[2]

As an example, in an antidepressant-versus-placebo RCT, we may assess the severity of illness using MADRS, HDRS, CGI-S, CGI-I, and other scales. We must decide a priori which one of these will be the primary outcome measure. We cannot decide after seeing the results which to project as the primary outcome, because that might well have been a chance finding.

The Primary Outcome Measure and Protection Against a Type II Error

A clinical trial that is conducted should answer the research question that was set; for example, whether a particular treatment is effective for a particular purpose. If the sample size is small and the study fails to show that the treatment is effective, then one of 2 explanations is possible:

1. The treatment is truly ineffective.
2. The treatment is effective, but the study failed to identify a statistically significant advantage because the sample size was too small. This is known as a false-negative or type II statistical error arising from insufficient statistical power.

Studies that fail because of inadequate sample size are unethical and wasteful (see sidebar). Therefore, investigators need to a priori estimate the minimum sample size required to answer a particular research question. This value can be calculated. However, the estimated sample size will vary depending on what the research question is. As already discussed, every clinical trial will have a number of efficacy and safety outcome measures; therefore, every clinical trial will also have a number of research questions, related to the many outcome measures. The necessary sample size to adequately power each research question would result in the silly situation of requiring many different sample sizes for the same study. The investigator cannot solve the problem by taking the largest value for sample size, because the really important questions, such as those related to efficacy of the experimental treatment, may be answered through a much smaller sample. To cut a long story short, the investigator chooses 1 safety or efficacy question that best justifies the trial, and he sets this as the primary outcome measure. The sample size is then estimated for this outcome. The study will then be adequately powered for the primary outcome but not necessarily for the secondary outcomes detailed in the plan of analysis. Thus, statistical testing of secondary outcomes may yield false-negative results.

Here are 2 examples. An investigator may define treatment efficacy as HDRS scores falling by at least 3 points more in the experimental antidepressant group as compared with the placebo group at the 8-week treatment endpoint in an intent-to-treat analysis. Or, successful outcome may be defined as the experimental antidepressant resulting in an at least 10% greater response rate than placebo. In each of these examples, the investigator could calculate the sample size necessary for him to be at least 80% certain of identifying a statistically significant result ($P < .05$) should the antidepressant drug be truly superior to placebo. This would be an adequately powered trial.

When the Primary Outcome Is Not Specified

Original research articles that do not specify a primary outcome have been published and continue to be published. What should a reader make of the results of these studies? In such articles, readers should consider that there is a higher than average likelihood that the article emphasizes results that support the objectives of the study, underplaying or omitting the inconvenient findings. As stated earlier, what the primary outcome was can be ascertained from the clinical trial registry if the article describes the results of a clinical trial and if the clinical trial was registered in the country in which it was conducted, or elsewhere. This information can be identified through a simple Internet search.

Reasons Why Clinical Trials That Are Inadequately Powered Are Unethical and Wasteful

Unethical:
Because the failure to reach a definite conclusion means that patients would have been inconvenienced or exposed to risk without benefit either to future patients or to the cause of science. For example, patients would have been inconvenienced by having to attend frequent study-related follow-ups, they would have been subjected to physical discomfort by blood draws for safety assessments, and they would have been exposed to a potentially ineffective treatment such as placebo, or even the experimental treatment if the treatment is actually not effective.

Wasteful:
Because the investigating team, the study sponsors or funding agency, the institutional review board or ethics committee, and others involved in the study would have expended much time, money, and effort to perform a study that neither benefited future patients nor furthered the cause of science.

Some studies are exploratory and have no primary outcome. Such studies are at increased risk of false-positive errors. Data mining and data dredging studies are particular examples of studies at high risk of false-positive findings.[5]

Outcomes That Are Neither Primary Nor Secondary

Primary and secondary outcomes are specified a priori in clinical trial protocols. Sometimes, investigators test hypotheses that are set a posteriori, that is, after discovering a pattern in the findings. The results of such hypothesis testing must be viewed with caution because the pattern may represent a chance finding and may not exist in other sets of data. Occasionally, a posteriori hypotheses may represent genuine results that were serendipitously discovered.[6–8] Until such results are confirmed prospectively or in other datasets, the results must be considered preliminary and not definitive. Some guidance is available for a posteriori testing of results in subgroups.[2]

Quandary

In most contexts in psychiatry, there is no gold standard for the choice of primary outcome. The primary outcome is then selected by expert consensus or by following conventional practice. After the study is completed and the data are analyzed, it may be found that, whereas the primary outcome for efficacy is not statistically significant, 1 or more secondary outcomes for efficacy are significant. This result can be interpreted in one of several ways:

1. The treatment is truly effective, and the investigator may have chosen an inappropriate primary outcome when the secondary outcomes were more important.
2. The treatment is truly ineffective, as shown by the results for the primary outcome, and the secondary outcomes were significant by chance.
3. The treatment is truly effective for some outcomes, such as the secondary outcomes, and truly ineffective for some outcomes, such as the primary outcome. This can happen when the primary and secondary outcomes assess different constructs, something that may not always be apparent.

There is no way of knowing for certain which interpretation is correct. Readers are reminded, for example, that lamotrigine separated from placebo on most secondary outcome measures but not on the primary outcome measure in the first trial in bipolar depression[9]; later, however, a meta-analysis showed that the drug is effective for this indication.[10]

It would be a travesty of research principles to regard secondary outcome measures as outcomes to fall back on in case the primary outcome measure fails to yield results that satisfy the investigators.[3]

Other Limitations of the Concept of Primary and Secondary Outcome Measures

In an antidepressant RCT, if the drug outperforms placebo on the primary outcome measure, it means that the drug is effective with regard to this measure. However, investigators and readers alike tend to equate efficacy on a primary outcome with efficacy across the board for that drug and indication; qualifiers are seldom specified. What this means is that efficacy on a clinician-rated instrument cannot be generalized to efficacy on a patient-rated instrument; efficacy on a depression rating scale cannot be generalized to efficacy on global ratings or ratings of quality of life; and efficacy in one environment (eg, at home, in children with attention-deficit/hyperactivity disorder) cannot be generalized to efficacy in all environments (eg, school and elsewhere). These comments apply to secondary outcomes as well as primary outcomes and to results demonstrating inefficacy as well as results demonstrating efficacy. To overcome these limitations, alternatives need to be developed; interested readers are referred to a technical discussion provided by De Los Reyes et al.[3]

Parting Notes

Although the concept of primary outcome measure has been discussed in this article in the context of clinical trials, the concept is applicable to observational studies, laboratory studies, and other kinds of original research, as well. The implications are the same.

Acknowledgment: Dr Andrade thanks Prof David Streiner, PhD, CPsych, Professor, Department of Psychiatry & Behavioural Neurosciences, McMaster University, and Professor, Department of Psychiatry, University of Toronto, for his careful reading of a previous version of this manuscript and his suggestions for its improvement.

REFERENCES

1. Sedgwick P. Primary and secondary outcome measures. *BMJ*. 2010;340:c1938.
2. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ*. 2001;322(7292):989–991.
3. De Los Reyes A, Kundey SM, Wang M. The end of the primary outcome measure: a research agenda for constructing its replacement. *Clin Psychol Rev*. 2011;31(5):829–838.
4. Vedula SS, Bero L, Scherer RW, et al. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *N Engl J Med*. 2009;361(20):1963–1971.
5. Andrade C. Antidepressants and testicular cancer: cause versus association. *J Clin Psychiatry*. 2014;75(3):e198–e200.
6. Kurinji S, Andrade C. ECS seizure threshold: normal variations, and kindling effects of subconvulsive stimuli. *J ECT*. 2003;19(1):31–37.
7. Andrade C, Akki A, Nandakumar N, et al. Confirmation of whole-brain kindling with repeated subthreshold electroconvulsive shocks: a controlled study. *J ECT*. 2003;19(2):81–83.
8. Andrade C. In vigorous defense of whole brain kindling and a reconsideration on naming the phenomenon. *J ECT*. 2004;20(4):275–276.
9. Calabrese JR, Bowden CL, Sachs GS, et al. A double-blind placebo-controlled study of lamotrigine monotherapy in outpatients with bipolar I depression. Lamictal 602 Study Group. *J Clin Psychiatry*. 1999;60(2):79–88.
10. Geddes JR, Calabrese JR, Goodwin GM. Lamotrigine for treatment of bipolar depression: independent meta-analysis and meta-regression of individual patient data from five randomised trials. *Br J Psychiatry*. 2009;194(1):4–9.