DOI: 10.5455/bcp.20110706020600

The Tyranny of the P-value: Effect Size Matters

Leslie Citrome[1]

ÖZET:

P değeri zorbalığı: Etki büyüklüğü sorunu

İstatistiksel önemlilik mutlaka klinik önemlilik anlamına gelmez. P-değerinin 0.05'ten az olması sonucun önemli olacağını garanti etmez. Kliniksel ilişkinin değerlendirebilmesi için etki büyüklüğü (effect size) hesaplanması gerekmektedir. Number needed to treat (NNT) klinik çalışma sonuçlarını anlamaya yardımcı olan ve bir klinisyenin rakip girişimler arasındaki potansiyel farkları değerlendirmesine izin veren örnek bir etki büyüklüğü ölçüsüdür.

Anahtar sözcükler: Klinik önemlilik, tedavi edilmesi gereken sayı, P-değeri, istatistiksel önemlilik

Klinik Psikofarmakoloji Bülteni 2011;21(2):91-2

ABSTRACT:

The tyranny of the P-value: effect size matters

Statistical significance does not necessarily imply clinical significance. A P-value of less than 0.05 does not guarantee that the result will be important. Effect size needs to be calculated in order to appraise clinical relevance. Number needed to treat (NNT) is an example of an effect size measure that helps translate clinical trial results and allows a clinician to evaluate potential differences between competing interventions.

Key words: Clinical significance, number needed to treat, P-value, statistical significance

Bulletin of Clinical Psychopharmacology 2011;21(2):91-2

[1]M.D., MPH, Professor of Psychiatry, New York University School of Medicine, New York-USA

Yazışma Adresi / Address reprint requests to: Leslie Citrome, MD, MPH, 11 Medical Park Drive, Suite 106, Pomona, NY 10970

Telefon / Phone: + 1 845 362 2081

Faks/ Fax: + 1 845 362 8745

Elektronik posta adresi / E-mail address: nntman@gmail.com

Kabul tarihi / Date of acceptance: 4 Nisan 2011 / April 4, 2011

Bağıntı beyanı:
L.C.: Yazar bu makale ile ilgili olarak herhangi bir çıkar çatışması bildirmemiştir.

Declaration of interest:
L.C.: The author reported no conflict of interest related to this article.

A common misconception is that in order for a clinical trial study result to be important, the outcome merely needs to be statistically significant. Obtaining a P-value of less than 0.05 becomes the goal in order to have a paper to be of interest to journal editors and deemed worthy of publication (1).

Unfortunately the "tyranny of the P" can lead researchers astray, allow irrelevant findings to be published, and confuse clinicians eager for guidance as to the usefulness of potential interventions. Take for example a hypothetical clinical trial of 2 medications for a major depressive episode; let's call them antidepressant A and B. Antidepressant A resulted in a superior rate of clinical response than antidepressant B. The P-value was less than 0.0001; highly statistically significant! However a closer look at this finding will reveal some additional information that renders A's superiority to B highly clinically irrelevant. This hypothetical study enrolled about 70,000 patients. Responder rates were 37.5% and 35.5% for antidepressants A and B, respectively. Not only are the response rates mediocre, but the difference in response rates was a mere 2%. Clearly, the P-value was not helpful in discerning clinical significance. In order to quantify the clinical importance of a statistically significant result, we need to calculate the effect size.

The effect size of a treatment represents how large a clinical response is observed (2-4). For continuous outcome measures like a point change on a rating scale, the effect size can be standardized so that it is easier to compare treatment effects in a meta-analysis. For categorical outcomes such as whether or not response or remission was achieved, proportions can be directly compared by simple subtraction to calculate the effect size difference. In our hypothetical case above, the difference was 2%. We can take this one step further to calculate the "number need to treat" (NNT), a number that is clinically intuitive and helps relate the effect size difference back to the realities of clinical practice.

NNT answers the question "how many patients do I need to treat with treatment A versus treatment B before I would expect to encounter one additional outcome of interest, such as a response?" It is calculated by taking the reciprocal of the difference in the rates of the outcome of interest, i.e. NNT = 1 / (Rate1-Rate2). For our hypothetical

example, NNT = 1 / .02 = 50. It would require treating 50 patients with antidepressant A instead of antidepressant B before expecting to encounter one additional responder. Together with the lackluster response rates to begin with, it would take quite a while before the clinician would notice a relevant efficacy difference between these two antidepressants based on response rates alone. Other factors become more important, such as the relative tolerability and safety of the medications being considered, as well as their cost, and the individual patient's past history of response (or non-response).

The NNT is independent of the P-value. It is possible to calculate a NNT from a statistically non-significant result, but it would be impossible to appropriately interpret it. NNT is often presented together with its "confidence interval" (CI) that can inform us about the precision of the NNT. For example a NNT of 6 with a 95% CI of 4-9 is a relatively precise estimate as the CI is "narrow." However a NNT of 11 with a 95% CI of 6-45 is less precise as the CI is "wide." A non-statistically significant result will have a NNT that has a CI that includes "infinity"; thus an infinite number of patients would need to be treated with intervention A versus intervention B before one additional outcome of interest would be expected, i.e. there is no difference. Formulas for the CI, as well as for other measures of effect size can be found elsewhere (5).

The smaller the NNT, the more important the effect size difference is between the two interventions you are comparing. Thus an NNT of 2 would mean you would expect to encounter an additional outcome of interest every 2 patients treated with one intervention versus the other. In general, "single-digit" NNTs would be clinically relevant because you would be encountering these differences in daily clinical practice. NNTs that are greater than 10 are generally less important, and those greater than 100 are largely irrelevant, unless the outcome is particularly serious, such as avoidance of myocardial infarction or death.

NNTs are by convention presented as whole numbers. NNT is a clinician's tool and clinicians do not treat fractions of patients. In order to be conservative and not exaggerate potential differences, NNT values are "rounded-up" to the next highest whole number. Thus a NNT of 5.3 gets rounded-up to 6.

When comparing interventions in terms of adverse outcomes we can refer to the NNT as a "number needed to harm" and is abbreviated NNH. The calculation of NNT and NNH is identical. Possible outcomes subject to this calculation include the occurrence of weight gain in excess of a certain threshold such as 7%, or the occurrence of akathisia, or a complaint of sedation (5,6).

NNT is not the perfect effect size measure. It is only calculable for binary or dichotomous outcomes at a specific point in time. NNT does not capture information about trajectory of improvement. When converting a continuous measure into a binary one, for example when taking a rating scale score and classifying those that achieve a certain degree of reduction as "responders," some information is lost.

Although an effect size measure such as NNT is only as good as the data behind it, it helps appraise the potential importance of a statistically significant result. I encourage all clinicians to look beyond the P, and ask the question "Is this result clinically relevant?"

References:

1. Citrome L. Call for papers for the International Journal of Clinical Practice. Video. Available at: http://www.youtube.com/watch?v=KBALRk2hjMs
2. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry 2006;59(11):990-6.
3. Citrome L. Compelling or irrelevant? Using number needed to treat can help decide. Acta Psychiatr Scand 2008;117(6):412-9.
4. Citrome L. Relative vs. absolute measures of benefit and risk: what's the difference? Acta Psychiatr Scand 2010;121(2):94-102.
5. Citrome L. Quantifying risk: The role of absolute and relative measures in interpreting risk of adverse reactions from product labels of antipsychotic medications. Curr Drug Saf 2009;4(3):229-37.
6. Citrome L. Adjunctive aripiprazole, olanzapine, or quetiapine for major depressive disorder: an analysis of number needed to treat, number needed to harm, and likelihood to be helped or harmed. Postgrad Med 2010;122(4):39-48.