JAMA Guide to Statistics and Methods

Sample Size Calculation for a Hypothesis Test

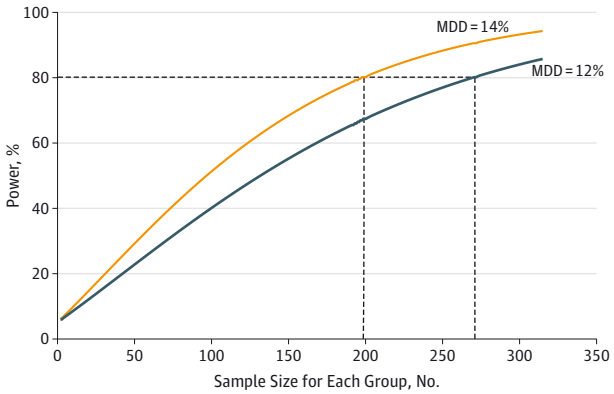
Lynne Stokes, PhD

In this issue of *JAMA*, Koegelenberg et al[1] report the results of a randomized clinical trial (RCT) that investigated whether treatment with a nicotine patch in addition to varenicline produced higher rates of smoking abstinence than varenicline alone. The primary results were positive; that is, patients receiving the combination therapy were more likely to achieve continuous abstinence at 12 weeks than patients receiving varenicline alone. The absolute difference in the abstinence rate was estimated to be approximately 14%, which was statistically significant at level α = .05.

Related article page 155

These findings differed from the results reported in 2 previous studies[2,3] of the same question, which detected no difference in treatments. What explains this difference? One explanation offered by the authors is that the previous studies "…may have been inadequately powered," which means the sample size in those studies may have been too small to identify a difference between the treatments tested.

Figure. Power for Detecting Difference and Sample Size



For a baseline rate of 45% and a minimum detectable difference (MDD) of 14%, the target sample size of 398 (199 in each group) will produce a power of 80% when α is set to .05. When the MDD is 12%, the resulting sample size is 542 (2 × 271) to achieve a power of 80%.

Use of the Method

Why Is Power Analysis Used?

The sample size in a research investigation should be large enough that differences occurring by chance are rare but should not be larger than necessary, to avoid waste of resources and to prevent exposure of research participants to risk associated with the interventions. With any study, but especially if the study sample size is very small, any difference in observed rates can happen by chance and thus cannot be considered statistically significant.

In developing the methods for a study, investigators conduct a power analysis to calculate sample size. The power of a hypothesis test is the probability of obtaining a statistically significant result when there is a true difference in treatments. For example, suppose, as Koegelenberg et al[1] did, that the smoking abstinence rate were 45% for varenicline alone and 14% larger, or 59%, for the combination regimen. Power is the probability that, under these conditions, the trial would detect a difference in rates large enough to be statistically significant at a certain level α (ie, α is the probability of a type I error, which occurs by rejecting a null hypothesis that is actually true).

Power can also be thought of as the probability of the complement of a type II error. If we accept a 20% type II error for a difference in rates of size *d*, we are saying that there is a 20% chance that we do not detect the difference between groups when the difference in their rates is *d*. The complement of this, 0.8 = 1 − 0.2, or the statistical power, means that when a difference of *d* exists, there is an 80% chance that our statistical test will detect it.

The **Figure** illustrates the relationship between sample size and power for the test described. The orange line shows the power for the parameter settings above (baseline rate of 45% and minimum detectable difference, or MDD, of 14%), when significance level α is set to .05. For this scenario, the authors' target sample size of 398 (199 in each group) will produce a power of 80%. All these values (45%, 14%, .05, 80%) must be selected at the planning stage of the study to carry out this calculation. The significance level and power are "rule-of-thumb" choices and are typically not based on the specifics of the study. If the researcher wants to reduce the probability of making a type I error (α = .05) or to increase the probability of detecting the specified difference (power = 80%), then these values can be changed. Either change will require a larger sample size.

Selecting the baseline rate and MDD requires the expertise of the researcher. The baseline rate is typically available from the literature, because this rate is often based on a therapy that has been studied. The MDD choice is more subjective. It should be a clinically meaningful rate difference, or a scientifically important rate difference, or both, that is also feasible to detect. For example, if the combination therapy of varenicline and nicotine patch increased abstinence by 0.1%, this difference would not be of practical benefit, would require an extremely large sample size, and would thus be too small a setting for the MDD. If the MDD were specified as 50%, the new therapy would have to be 95% effective (45% + 50%) before there would be a high chance of detecting any difference, so would be too large for the MDD. The authors based their choice of MDD = 14% on a compromise between their judgment of a clinically important difference, 12%, and the scientifically meaningful value of 16%. The 16% rate was the observed difference in a study that compared varenicline alone and together with nicotine gum.[4] Thus, the ability to con-

firm a difference that is slightly smaller for a related treatment was considered scientifically important.

What Are the Limitations of Power Analysis?

Calculation of sample size requires predictions of baseline rates and MDD, which may not be readily available, before the study begins. The sample size is especially sensitive to the MDD. This is illustrated by the blue line in the Figure, which shows the sample size needed in this study if the MDD were set to 12%. The resulting sample size is 542 (2 × 271) to achieve a power of 80%.

This method of conducting a power analysis might also produce the incorrect sample size if the data analysis conducted differs from that planned. For example, if abstinence were affected by other covariates, such as age, and the groups were unbalanced on this variable, other analyses might be used, such as logistic regression models accounting for covariate differences. The sample size that would be appropriate for one analysis may be too large or small to achieve the same power with another analytic procedure.

Why Did the Authors Use Power Analysis in This Particular Study?

The number of research participants available for any study is limited by resources. However, the authors were aware that previous studies comparing these treatments had found no significant difference in abstinence rates. This can occur even if a difference exists if the sample size is too small. The authors wanted to ensure that their sample size was adequate to detect even a small but clinically important difference, so they carefully evaluated sample size.

How Should This Method's Findings Be Interpreted in This Particular Study?

A power analysis can help with the interpretation of study findings when statistically significant effects are not found. However, because the findings in the study by Koegelenberg et al[1] were statistically significant, interpretation of a lack of significance was unnecessary. If no statistically significant difference in abstinence rates had been found, the authors could have noted that, "The study was sufficiently powered to have a high chance of detecting a difference of 14% in abstinence rates. Thus, any undetected difference is likely to be of little clinical benefit."

Caveats to Consider When Looking at Results Based on Power Analysis

Sample size calculation based on any power analysis requires input from the researcher prior to the study. Some of these are assumptions and predictions of fact (such as the baseline rate), which may be incorrect. Others reflect the clinical judgment of the researcher (eg, MDD), with which the reader may disagree. If a statistically significant effect is not found, the reader should assess whether either of these are concerns.

The reader should also not interpret a lack of significance for an outcome other than the one on which the power analysis was based as confirmation that no difference exists, because the analysis is specific to the parameter settings. For example, no significant difference was found in this study for most adverse events rates, although the power analysis does not apply to these rates. Thus, the sample size may not be adequate to interpret that finding to confirm that no meaningful difference in these outcomes exists.

ARTICLE INFORMATION

Author Affiliation: Department of Statistical Science, Southern Methodist University, Dallas, Texas.

Corresponding Author: Lynne Stokes, PhD, Department of Statistical Science, Southern Methodist University, PO Box 750100, Dallas, TX 75275 (slstokes@smu.edu).

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Koegelenberg CFN, Noor F, Bateman ED, et al. Efficacy of varenicline combined with nicotine replacement therapy vs varenicline alone for smoking cessation: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2014.7195.

2. Hajek P, Smith KM, Dhanji AR, McRobbie H. Is a combination of varenicline and nicotine patch more effective in helping smokers quit than varenicline alone? a randomised controlled trial. *BMC Med*. 2013;11:140.

3. Ebbert JO, Burke MV, Hays JT, Hurt RD. Combination treatment with varenicline and nicotine replacement therapy. *Nicotine Tob Res*. 2009;11(5):572-576.

4. Besada NA, Guerrero AC, Fernandez MI, Ulibarri MM, Jiménez-Ruiz CA. Clinical experience from a smokers clinic combining varenicline and nicotine gum. *Eur Respir J*. 2010;36(suppl 54):462s.