JACC: BASIC TO TRANSLATIONAL SCIENCE VOL. 9, NO. 2, 2024

EDITOR'S PAGE

Is it Time to Abandon the Use of *P*-Values in Early Phase Translational Trials: Why (Effect) Size Matters

Douglas L. Mann, MD

Phase 1 clinical trials are primarily designed to test the safety of new drugs or devices. However, increasingly, phase 1 studies are being used as hypothesis-generating studies to help inform endpoints for larger phase 2 studies. Traditionally, researchers have relied on assessing a number of different possible endpoints using *P* values as the primary way of determining the significance of their findings. However, the use of *P* values in small clinical studies is fraught for several reasons, not the least of which is that if a study is underpowered to detect a real difference (ie, low prestudy odds), then there is <25% change that the positive research finding is actually true and is, therefore, unlikely to be replicated in phase 2 to 3 clinical trials.[1,2]

Effect size is a simple, albeit underutilized, method for assessing the clinical significance of research findings in small clinical trials that are underpowered to see statistically significant differences in trial endpoints. Effect size is a quantitative measure of the difference between 2 groups, and can be measured in a number of ways, including the standardized mean difference (SMD) for comparing the means between 2 groups, the Pearson's r for measuring the strength and direction of a linear relationship between 2 continuous variables, the odds ratio for comparing the odds of an outcome occurring in 2 groups (common in case-control studies), and the relative risk for comparing the risk of an outcome in 2 groups (common in cohort studies).

Among these different methodologies, the SMD is perhaps the easiest method to utilize in small clinical studies that evaluate a variety of different exploratory endpoints. The SMD can be calculated as (mean$^{\text{study drug/device}}$ − mean $^{\text{control arm}}$)/standard deviation (SD). An SMD of 0 indicates that the drug/device has equivalent effects to those observed in the control arm, whereas a value >0 indicates that the drug/device has a beneficial effect when compared with control values, and a value <0 indicates that a drug/device is worse than no intervention. The inclusion of the SD in the denominator of the equation adjusts for the variability in the measurements in the drug/device and no treatment arms, and standardizes the comparisons of the magnitude of treatment effects, which can be useful when evaluating multiple different exploratory endpoints. In addition to the simplicity of calculating the SMD, there are 3 additional aspects of this methodology that are extremely useful for interpreting the results of exploratory analyses in small phase 1 clinical studies. The first is that the use of effect sizes allows investigators to address the question of whether the magnitude of change between the treatment and control groups is important and clinically meaningful. There are several published guidelines for interpreting effect sizes, of which Cohen's *d* is the most widely known.[3] Cohen suggested a convention to interpret the magnitude of the effect size, where $d = 0.2$ is a small effect, $d = 0.5$ is a moderate effect, and $d = 0.8$ is a large effect. Cohen originally proposed that a medium-sized effect should represent the average effect size within the field (ie, 50th percentile), with a small effect size associated with the 25th percentile and a large effect size reflecting the 75th percentile. Not surprisingly, subsequent studies have shown that the magnitude of small, medium, and large effects sizes varies from field-to-field, and that Cohen's *d* can overestimate or underestimate the actual effect size distribution in the published literature. Although a distribution of

ISSN 2452-302X https://doi.org/10.1016/j.jacbts.2024.01.008

actual effect sizes has not been assessed for the entire field of cardiovascular medicine, a study that reported on effect sizes and primary outcomes in large cardiovascular-related behavioral randomized clinical trials revealed that effect sizes of the behavioral and physiological outcomes were predominantly in the small ($d = 0.2$) to medium ($d = 0.5$) range.[4] However, it should be recognized that clinical significance of an absolute effect size is context-dependent. For example, a small effect size for mortality can make a huge difference for society if a large proportion of the population is affected by the condition (eg, the COVID-19 pandemic). Another useful aspect of reporting effect sizes in small clinical trials is that one can calculate confidence intervals (CIs) around the point estimates for the effect size, which provides a range of plausible values for the variable being measured. There is a close relationship between CIs and statistical significance testing. If the 95% CI for the point estimate of the effect size contains a value that contains the null value of 0, the difference will be nonsignificant if $P < 0.05$ is used as the cut-point for determining statistical significance. CIs can also be particularly useful in assessing multiple endpoints in exploratory analyses where statistical methodology for multiplicity of testing has not been implemented. Relevant to the present discussion, if the CIs encompass clinically meaningful differences for a given endpoint but the small sample size yields a nonsignificant result, this can also be informative insofar as it suggests that magnitude of change in the endpoint may be clinically relevant. Last, effect sizes can also be used to perform power calculations for larger phase 2 to 3 clinical trials.

Despite the importance of measuring and reporting effect sizes in small clinical trials, there are several caveats that warrant discussion. As noted by Ioannidis,[5] newly discovered findings may report inflated effect sizes when compared with the true effect sizes if the discovery is based on crossing a threshold of statistical significance, and the discovery study is underpowered to observe statistically significant differences. This is particularly important in small clinical trials where multiplicity of testing can lead to type I statistical errors. Second, selective reporting of endpoints, which is common in smaller clinical trials where the goal is to conduct multiple exploratory analyses to discover something new, can be problematic if the endpoints selected for presentation are the largest effect sizes. Last, studies have shown that estimates for treatment effect sizes are significantly larger in small clinical studies than in larger clinical trials, which may be related to wider eligibility criteria used in larger trials.[6]

As noted previously in these Editor's Pages,[7] early-phase translational research studies are inherently fragile because of the small sample sizes that are employed. In translational studies that employ multiple exploratory analyses, it is often difficult to strike the correct balance between discovery and replicability. Despite all of the widely recognized limitations of statistical significance testing in small clinical studies, it is unlikely that *P* values are going away anytime soon because they are so deeply embedded in the statistical culture of how we evaluate clinical research studies. Assessing effect size and CIs in small clinical studies may allow translational investigators to overcome some of the inherent problems with *P*-hacking and multiplicity of statistical testing that are rife in translation research studies. However, as noted, the assessment of effect size is not without limitations and caveats. Although these arguments are certainly not new, they have traditionally been discussed in the context of large phase 3 clinical trials. Here, we suggest that understanding and appropriately interpreting effect size along with formal statistical testing is equally important for translational scientists who are called upon to make go/no go decisions about whether or not advance a new drug/device into larger phase 2 to 3 clinical trials.

ADDRESS FOR CORRESPONDENCE: Dr Douglas L. Mann, *Editor-in-Chief, JACC: Basic to Translational Science,* American College of Cardiology, Heart House, 2400 N Street Northwest, Washington, DC 20037, USA. E-mail: JACCBTS@acc.org.

REFERENCES

1. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.

2. Kraemer HC. The average error of a learning model, estimation and use in testing the fit of models. *Psychometrika*. 1965;30:343–352.

3. Cohen J. *The Concept of Power Analysis*. Routledge; 1988.

4. Irvin VL, Kaplan RM. Effect sizes and primary outcomes in large-budget, cardiovascular-related behavioral randomized controlled trials funded by National Institutes of Health since 1980. *Ann Behav Med*. 2016;50:130–146.

5. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. 2008;19:640–648.

6. Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *BMJ*. 2013;346:f2304.

7. Mann DL. Deus ex machina: why mechanism matters in translational research. *J Am Coll Cardiol Basic Trans Science*. 2017;2:227–228.