MedCrave
Step into the World of Research
Biometrics & Biostatistics International Journal

Review Article
Open Access
CrossMark

Statistical significance vs. clinical significanc

Abstract

When designing a clinical study to compare the efficacy of an experimental treatment with a control, the primary endpoint is selected to measure the treatment effect, along with a clinically meaningful effect defined for that endpoint. At the end of the study, success is typically determined by statistical significance, based on a test statistic and its associated p-value calculated from the observed data. In addition to statistical testing, the observed treatment effect is compared with the pre-specified clinically meaningful threshold to assess clinical significance. The sample size is calculated to ensure high statistical power that is, a probability of detecting statistical significance for the specified clinically meaningful effect. However, the probability of achieving clinical significance is generally lower than the statistical power. As a result, it is not uncommon for a study to demonstrate statistical significance without clinical significance. In this article, we examine the relationship between clinical and statistical significance under various scenarios. This analysis helps explain why statistically significant results frequently lack clinical significance. If we seek a higher probability of clinical significance, our studies must be designed accordingly.

Keywords: clinically meaningful effect size, primary endpoint, strength of evidence, sample size calculation, standardized effect size

Volume 14 Issue 2 - 2025

Chaewon Jeong, Sin Ho Jung
Department of Biostatistics and Bioinformatics, Duke University, USA

Correspondence: Sin Ho Jung, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA, Tel 919-668-8658

Received: September 15, 2025 | **Published:** December 23, 2025

Introduction

The relationship between statistical significance and clinical significance has been a prominent topic in the interpretation of clinical research results.[1–6] When designing a clinical study to compare the efficacy of an experimental treatment with that of a control, a primary endpoint is selected to measure the treatment effect, along with a clinically meaningful effect defined in terms of that endpoint. At the end of the study, the success of the trial is typically determined by statistical significance, based on a pre-specified test statistic and its associated p-value calculated from the observed data. In addition to statistical testing, the observed treatment effect is also compared with the pre-specified clinically meaningful threshold to assess clinical significance. Since the sample size for a study is calculated to detect the specified clinically meaningful effect with a given level of statistical power, it is possible for a study to demonstrate statistical significance without clinical significance.

A study is declared positive if it achieves statistical significance. Some investigators complain that it has not achieved clinical significance. Page[7] discusses the distinction between clinical and statistical significance, emphasizing the use of confidence intervals (CIs) rather than statistical tests to express clinical relevance. However, the roles of CIs and statistical tests are essentially equivalent, as both lead to the same conclusions regarding statistical and clinical significance. Moreover, Page defines clinical significance in terms of all efficacy and safety outcomes of the two treatments, whereas in this article we focus specifically on the primary outcome that determines study success. Sharma[6] and Fethney[8] address the difference between clinical and statistical significance, with particular attention to the interpretation and role of p-values and confidence intervals, respectively. However, a p-value and confidence interval are directly related to statistical tests which determine statistical significance, so that it is difficult to associate them with clinical significance.

In this paper, the concept of clinical significance is based on statistical interpretation like statistical significance as in Elasan.[9] Based on this concept, we observe statistical significance and clinical insignificance more often than statistical insignificance and clinical significance from real clinical studies. We investigate why this happens, and when the probability of clinical significance can be increased.

The methods discussed vary slightly depending on the type of study endpoint. In this paper, we focus on studies with continuous and time to event endpoints, although the results can be readily extended to other types of endpoints like binary endpoints.

Materials and methods

We apply the t-test for continuous outcomes and the maximum likelihood test for time to event outcomes. Large sample approximations are used to evaluate the relationship between clinical significance and statistical significance under these testing methods. For various scenarios, we derive the probabilities of achieving clinical and statistical significance under both the null and alternative hypotheses. All calculations are implemented using FORTRAN programs developed by the authors.

Results

We investigate the relationship between clinical significance and statistical significance when the primary outcome is either continuous or a time-to-event endpoint.

Continuous endpoint

Suppose that the primary endpoint of a study is continuous, e.g. the change in tumor size before and after treatment in cancer patients. We compare the efficacy of an experimental treatment (group 1) with that of a standard treatment (group 2). The study is designed to test the null hypothesis that the experimental treatment is no more effective than the standard treatment, against the alternative hypothesis that the experimental treatment provides superior efficacy. Without loss of generality, we assume that the outcome data or its suitable transformation for group $k(=1,2)$ is normally distributed with mean μ_k and variance σ^2.

Statistical significance and clinical significance

In a clinical research with a continuous endpoint, the efficacy of two treatments is measured by the difference in mean $\mu_1 - \mu_2$ of the continuous endpoint. Clinical significance of the research is determined by the observed treatment effect from the resulting data. By contrast, statistical significance depends not only on the observed treatment effect but also on the strength of the supporting evidence, which is measured by the standard error (SE) of the estimate. A study result is clinically significant but not statistically significant if the observed treatment effect is meaningful yet the evidence is weak, i.e., SE is large. Conversely, it may be statistically significant but not clinically significant if the observed effect is small but the evidence is strong (i.e., SE is very small).

The null and alternative hypotheses are expressed as $H_0 : \mu_1 \le \mu_2$ and $H_1 : \mu_1 > \mu_2$, respectively. In the study, n_k patients are assigned to treatment group k and μ_k is estimated by sample mean $\bar{x}_k$.

Sharma[6] argues that clinical significance should be defined based on all relevant outcomes, including efficacy, safety, and quality of life. While this may be appropriate for a broad, general definition of clinical significance, when contrasted with statistical significance its definition should be narrowed to focus specifically on the pre-specified treatment effect for the primary endpoint.

At the design stage, we specify a clinically meaningful treatment effect δ_1. For the prespecified clinically meaningful treatment effect, the study demonstrates clinical significance if the observed treatment effect, $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ exceeds δ_1.

The strength of the evidence refers to how consistent the observed treatment effect is across patients in the study. This is quantified by SE of the observed treatment effect, $\mathrm{SE}(\bar{x}_1 - \bar{x}_2) = \hat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}$, where $\hat{\sigma}^2$ is the estimator of $\hat{\sigma}^2$ from the resulting data. Thus, the strength of the evidence depends on three factors: the variance of the observations, the total sample size, and the allocation proportions between the two treatment groups.

More specifically, an evidence-adjusted treatment effect is expressed by the test statistic:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma} / \sqrt{n a_1 a_2}}$$

Where $n = n_1 + n_2$ the total is sample size and $a_k = n_k / n$ is the allocation proportion for group k $(a_1 + a_2 = 1)$. With a false positivity (type I error) rate α level, usually 0.05 or 0.025, we obtain statistical significance if z is larger than $z_{1-\alpha}$, the $100(1-\alpha)$ percentile of the standard normal distribution, or equivalently if

$$\bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n a_1 a_2}} \qquad (1)$$

Sample size calculation for statistical significance

As discussed in the previous section, statistical significance depends on the sample size, whereas clinical significance does not. Consequently, whether statistical and clinical significance align depends not only on the sample size but also on the variance of the observations. This issue is therefore closely tied to sample size determination during study design. In this section, we examine the relationship between clinical and statistical significance in the context of sample size calculations aimed at achieving statistical significance.

When designing a clinical study, the sample size is calculated to achieve statistical significance with a high probability $1-\beta$, known as the study's statistical power, when the true treatment effect $\delta = \mu_1 - \mu_2$ equals the pre-specified clinically meaningful effect δ_1. The required total sample size is given as[10]

$$n = \frac{\sigma^2 (z_{1-\alpha} + z_{1-\beta})^2}{a_1 a_2 \delta^2} \qquad (2)$$

By substituting (2) into equation (1) and replacing $\hat{\sigma}^2$ with σ^2 (since $\hat{\sigma}^2$ is a consistent and unbiased estimator of σ^2), we conclude that statistical significance is achieved if

$$\bar{x}_1 - \bar{x}_2 > \frac{z_{1-\alpha}}{z_{1-\alpha} + z_{1-\beta}}\delta_1 \qquad (3)$$

Thus, using the calculated sample size, we will declare statistical significance if the observed treatment effect $\bar{x}_1 - \bar{x}_2$ exceeds a fraction $z_{1-\alpha} / (z_{1-\alpha} + z_{1-\beta})$ of the pre-specified clinically meaningful treatment effect δ_1. Since this fraction lies between 0 and 1, the probability of achieving statistical significance is always higher than that of achieving clinical significance when σ^2 is the true variance value and the clinically meaningful treatment effect δ_1 equals the true treatment effect $\delta = \mu_1 - \mu_2$.

If we assume $\alpha = \beta$ in a sample size calculation, then, from (3), we would conclude statistical significance if the observed treatment effect $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ exceeds half of the clinically meaningful effect δ_1. In practice, however, we usually assume $\alpha < \beta$, so that $z_{1-\alpha} > z_{1-\beta}$, which makes the fraction larger than half. For example, with the commonly used values $\alpha = 0.025$ and $1-\beta = 0.8$, we have

$$\frac{z_{1-\alpha}}{z_{1-\alpha} + z_{1-\beta}} = \frac{1.96}{1.96 + 0.842} = 0.7$$

In this case, a statistical significance will be declared if $\bar{x}_1 - \bar{x}_2 > 0.7\delta$, which can still be clinically insignificant if the observed effect is between $0.7\delta_1$ and δ_1.

On the other hand, if δ_1 equals the true treatment effect, then the probability of achieving clinical significance is 50% because

$$P(\bar{x}_1 - \bar{x}_2 > \delta_1 \mid \mu_1 - \mu_2 = \delta_1) = 0.5$$

Regardless of the sample size, allocation proportion, and variance σ^2. This is much smaller than the usual power level for statistical significance, so that we often observe clinical insignificance in spite of statistical significance.

Appendix A shows that, when the sample size is calculated using the correct variance value σ^2 and a clinically meaningful treatment effect δ_1, then the probability of observing statistical significance without clinical significance is $0.5 - \beta$, whereas the probability of observing statistical insignificance with clinical significance is 0. For example, if the sample size is calculated for a power of $1-\beta = 80\%$, then there is a 30% $(= 0.5 - (1 - 0.8))$ of chance of observing statistical significance without clinical significance.

The two decision rules for significance have different false positivity (type I error) rate too. For the decision rule of clinical significance, the false positivity under $H_0 : \mu_1 = \mu_2$ is

$$P(\bar{x}_1 - \bar{x}_2 > \delta_1 \mid \mu_1 = \mu_2) = P\left(\frac{\bar{x}_1 - \bar{x}_2}{\sigma / \sqrt{n a_1 a_2}} > \frac{\delta_1}{\sigma / \sqrt{n a_1 a_2}} \mid \mu_1 = \mu_2\right)$$

$$P(Z > \frac{\delta_1}{\sigma / \sqrt{na_1a_2}}) \quad (4)$$

Where Z is a $N(0,1)$ random variable in (4),

$$\frac{\delta}{\sigma / \sqrt{na_1a_2}} = z_{1-\alpha} + z_{1-\beta}$$

From (3), the false positivity of the decision rule for clinical significance is given as $\bar{\Phi}(z_{1-\alpha} + z_{1-\beta})$, which is much smaller than the false positivity for statistical significance $\alpha = \bar{\Phi}(z_{1-\alpha})$. Here, $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$ and $\bar{\Phi}(\cdot)$ denotes the cumulative density function of $N(0,1)$.

Is it possible to have clinical significance but statistical insignificance?

In the previous section, we observe that, if the clinica'lly meaningful treatment effect δ_1 equals the true treatment effect, $\delta = \mu_1 - \mu_2$ the probability of achieving clinical significance is fixed at $1/2$ regardless of the sample size, allocation proportion, and variance σ^2. Furthermore, it is impossible to have clinical significance and statistical insignificance if the sample size for statistical significance is calculated based on these parameter values.

In a real clinical research, we can have clinical significance and statistical insignificance if the study is so underpowered that the true power for statistical significance is smaller than the true positivity of clinical significance $1/2$. This typically happens when the sample size is calculated based on a variance value of $\bar{\sigma}^2$ that is much smaller than the true variance σ^2. Now the question is how small the specified $\bar{\sigma}^2$ should be relative to the true value σ^2. In this section, we assume that the clinically meaningful treatment effect δ_1 equals the true treatment effect $\delta = \mu_1 - \mu_2$.

Suppose the sample size is calculated based on a misspecified variance $\hat{\sigma}^2$ that is smaller than the true variance σ^2. In this case, $\hat{\sigma}^2$ will converge to $\bar{\sigma}^2$, so that, from (2) we have

$$\sqrt{na_1a_2} = (z_{1-\alpha} + z_{1-\beta})\sigma / \delta_1 \quad (5)$$

and statistical significance is determined by

$$\bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\frac{\sigma}{\sqrt{na_1a_2}} = \frac{z_{1-\alpha}\sigma}{(z_{1-\alpha} + z_{1-\beta})\bar{\sigma}}\delta_1 \quad (6)$$

By incorporating (1) with (5).

On the right hand side of (6), if $\sigma / \bar{\sigma}$ is larger than $(z_{1-\alpha} + z_{1-\beta}) / z_{1-\alpha}$, then the right hand side of (6) exceeds δ_1, and it is possible to observe clinical significance but statistical insignificance with probability

$$P\left\{\delta_1 < \bar{x}_1 - \bar{x}_2 \left\langle \frac{z_{1-\alpha}\sigma}{(z_{1-\alpha} + z_{1-\beta})\bar{\sigma}}\delta_1 \middle| \mu_1 - \mu_2 = \delta_1, \sigma^2 \right.\right\}$$

$$= \Phi\left\{z_{1-\alpha} - \frac{\bar{\sigma}}{\sigma}(z_{1-\alpha} + z_{1-\beta})\right\} - 1/2$$

Since, from (6), $\sqrt{na_1a_2}(\bar{x}_1 - \bar{x}_2 - \delta_1)/\sigma = \bar{\sigma}(z_{1-\alpha} + z_{1-\beta})(x_1 - \bar{x}_2 - \delta_1)/(\delta_1\sigma)$ is $N(0,1)$ forgiven (δ_1, σ^2). Hence, if $\bar{\sigma}/\sigma$ is smaller than $z_{1-\alpha}/(z_{1-\alpha} + z_{1-\beta})$, then this probability is positive. Note that this probability depends only on the ratio $\bar{\sigma}/\sigma$ and is independent of δ_1.

For example, suppose a study's sample size is calculated using $(\alpha, 1-\beta) = (0.025, 0.8)$ and (δ, σ^2), but the assumed variance for sample size calculation $\bar{\sigma}^2$, is only a half of the true variance σ^2. Then, the probability of observing clinical significance but statistical insignificance is

$$\Phi\{1.96 - (1.96 + 0.842)/2\} - 1/2 = \Phi(0.559) - 1/2 = 0.712 - 0.5 = 0.212$$

.Thus, there is roughly a 21% chance that the study shows a clinically meaningful effect but fails to reach statistical significance due to the underestimated variance.

In summary, we can have clinical significance and statistical insignificance only when the sample size is substantially underpowered.

When can we have the probability of clinical significance larger than $1/2$?

As shown in Section 3.1.2, the probability of achieving clinical significance is 0.5 if the specified clinically meaningful treatment effect δ_1, is identical to the true treatment effect. This is much smaller than the usual power level of 0.8 to 0.9 for statistical significance, so that we often observe clinical insignificance in spite of statistical significance.

In a real clinical study, the probability of clinical significance can be higher than $1/2$ if the true treatment effect $\delta = \mu_1 - \mu_2$ is larger than the clinically meaningful treatment effect δ_1. In this case, the question is how large the true treatment effect should be for a probability of clinical significance $1-\pi$ to exceed $1/2$ .Suppose that the sample size n is calculated by (2) using parameters $(\alpha, 1-\beta, a_1, \sigma^2)$, and clinically meaningful treatment effect δ_1.

If we desire the probability of clinical significance to be $1-\pi$ for $\pi < 0.5$, then, from (A.4) of Appendix B, the true treatment effect should be as large as

$$\delta = \delta_1\left(1 + \frac{z_{1-\pi}}{z_{1-\alpha} + z_{1-\beta}}\right)$$

Note that δ is larger than δ_1 in this equation.

For example, if $\alpha = 0.025$, $(z_{1-\alpha} = 1.96)$ and $1-\pi = 1-\beta = 0.8$, $(z_{1-\pi} = z_{1-\beta} = 0.842)$, then we have $\delta = 1.3\delta_1$. That is, for $(\alpha, 1-\beta) = (0.025, 0.8)$, if the sample size is calculated for a clinically meaningful treatment effect δ_1 and we want the probability of achieving clinical significance to match the statistical power for statistical significance, then the true treatment effect must exceed the pre-specified clinically meaningful effect by 30%.

If the true effect size δ is larger than the clinically meaningful treatment δ, however, the true power for statistical significance will increase too.

Time to event endpoint

Suppose that the primary endpoint of a study is time to event variable, such as overall survival in phase III cancer clinical trials. Let $\Lambda_k(t)$ denote the cumulative hazard rate of group k. In this section, we assume that group 1 is control and groups 2 is experimental. Using a time to event endpoint, treatment effect of two patient groups is usually expressed as a hazard ratio (HR) $\Delta = \Lambda_1(t)/\Lambda_2(t)$.

The log-rank test[11] has been a standard choice for testing $H_0 : \Delta = 1$ against $H_1 : \Delta > 1$. Earlier, however, George and Desu[12] consider a test statistic based on maximum likelihood estimators (MLEs) of exponential distributions and show that its statistical power closely matches with that of the log-rank test using simulations. Schoenfeld[13] and Kwak and Jung[14] derived sample size formulas for the log-rank test. These sample size formulas are approximately the same as that of the MLE test of George and Desu[12] under a nearby alternative hypothesis.

If two test statistics have comparable power, they are effectively interchangeable. Based on this reasoning, we use the MLE test, explicitly expressed in terms of the HR, to discuss the relationship between statistical and clinical significance.

Statistical significance and clinical significance

Suppose that group k has an exponential distribution with hazard rate λ_k. Then the HR is expressed as $\Delta = \lambda_1 / \lambda_2$. The MLE of λ_k is given as $\hat{\lambda}_k = D_k / X_k$, where D_k denotes the number of events and X_k denotes the total study time.[15]

At the design stage, we specify a clinically meaningful treatment effect Δ_1. For a prespecified clinically meaningful treatment effect Δ_1, the study demonstrates clinical significance if the observed treatment effect, $\hat{\Delta} = \hat{\lambda}_1 / \hat{\lambda}_2$ exceeds Δ_1.

By the MLE theory, the estimated log-HR $\log\hat{\Delta} = \log\hat{\lambda}_1 / \hat{\lambda}_2$ is approximately normal with mean $\log\Delta$ and variance $1/D_1 + 1/D_2$. Hence, an evidence-adjusted treatment effect is expressed by the test statistic:

$$Z = \frac{\log\hat{\Delta}}{\sqrt{1/D_1 + 1/D_2}}$$

With a type I error rate α, we obtain statistical significance if Z, is larger than $z_{1-\alpha}$, or equivalently if

$$\log\hat{\Delta} > z_{1-\alpha}\sqrt{1/D_1 + 1/D_2} \qquad (7)$$

Sample size calculation

Statistical power of statistical tests for censored survival data depends on number of events rather than number of patients. The required number of events for the MLE test (7) to detect a clinically meaningful treatment effect Δ_1, with a power of $1-\beta$ is given by

$$\left(\frac{\log\Delta_1}{z_{1-\alpha} + z_{1-\beta}}\right)^2 = D_1^{-1} + D2^{-1} \qquad (8)$$

by George and Desu[12] and Rubinstein et al.[16]

By substituting (8) into (7), with a sample size calculation, we obtain statistical significance if

$$\log\hat{\Delta} > \frac{z_{1-\alpha}}{z_{1-\alpha} + z_{1-\beta}}\log\Delta_1 \qquad (9)$$

That is, using the calculated sample size, we will declare statistical significance if the observed treatment effect $\log\hat{\Delta}$ exceeds a fraction $z_{1-\alpha} / \left(z_{1-\alpha} + z_{1-\beta}\right)$ of the pre-specified clinically meaningful log-treatment effect $\log\Delta_1$. Since this fraction lies between 0 and 1, the probability of achieving statistical significance is always higher than that of achieving clinical significance and we cannot have clinical significance and statistical insignificance if the clinically meaningful treatment effect Δ_1 equals the true treatment effect Δ.

On the other hand, if the clinically meaningful treatment effect Δ_1 equals the true treatment effect Δ, the probability of achieving clinical significance is

$$P(\hat{\Delta} > \Delta_1 \mid \Delta_1) = 0.5 \qquad (10)$$

Regardless of the sample size and allocation proportion. This is much smaller than the usual power level for statistical significance, so that we often observe clinical significance in spite of statistical significance.

As in the continuous endpoint case, if the sample size is calculated for the statistical test to detect a clinically meaningful treatment effect δ with power $1-\beta$, then the probability of observing statistical significance without clinical significance is $0.5-\beta$, and the probability of observing statistical insignificance with clinical significance is 0.

Is it possible to have clinical significance but statistical insignificance?

If the clinically meaningful treatment effect Δ_1 equals the true treatment effect Δ, by (10), the probability of achieving clinical significance is fixed at $1/2$ regardless of the sample size and allocation proportions. Furthermore, it is impossible to have clinical significance and statistical insignificance if the sample size for statistical significance is calculated based on this assumption, i.e. $\Delta_1 = \Delta$.

A real clinical research, however, can have clinical significance and statistical insignificance if it is so underpowered that the true power for statistical significance is smaller than $1/2$. Typically, the MLE test can be underpowered if its required number of events is calculated by assuming much larger hazard rates $\bar{\lambda}_k$ than the true hazard rate $\bar{\lambda}_k$, while the clinically meaningful treatment $\Delta_1 = \lambda_1 / \lambda_2 = \bar{\lambda}_1 / \bar{\lambda}_2$ is unchanged.

Let $\bar{D}_k$ denote the expected number of events if the hazard rate is $\bar{\lambda}_k$. Then the assumed variance is $\bar{\sigma}^2 = 1/\bar{D}_1 + 1/\bar{D}_2$ which is smaller than the true variance $\sigma^2 = 1/D_1 + 1/D_2$. Hence, this is identical to assuming a variance value much smaller than the true one in continuous endpoint case, so that we can possibly observe clinical significance but statistical insignificance.

When can we have the probability of clinical significance larger than 50%?

As shown in (10), the probability of achieving clinical significance is $1/2$ if the specified clinically meaningful treatment effect Δ_1, is identical to the true treatment effect. This is much smaller than the usual power level for statistical significance, so that we often observe clinical insignificance in spite of statistical significance.

In a real clinical study, the probability of clinical significance can be higher than $1/2$ if the true treatment effect $\Delta = \lambda_1 / \lambda_2$ is larger than the clinically meaningful treatment effect Δ_1 as in the continuous endpoint case. For exponential distributions, the variance depends on the treatment effect, but we assume that the variance $\sigma^2 = 1/D_1 + 1/D_2$ remains approximately unchanged. Under this assumption, the probability of clinical significance is

$$1 - \pi = P\left(\log\hat{\Delta} > \log\Delta_1 \mid \Delta\right)$$

$$= P\left(\frac{\log\hat{\Delta} - \log\Delta}{\sigma} > \frac{\log\Delta_1 - \log\Delta}{\sigma} \mid \Delta\right)$$

$$= \bar{\Phi}\left(\frac{\log\Delta_1 - \log\Delta}{\sigma}\right)$$

This probability will be larger than $1/2$ if $\Delta_1 < \Delta$.

Discussion

Our discussions have focused on continuous and time to events endpoints, but similar principles apply to other types of endpoints. For a continuous endpoint, the relationship between clinical significance and statistical significance depends on whether the true treatment effect, $\delta = \mu_1 - \mu_2$, exceeds the pre-specified clinically meaningful treatment effect and whether the variance, σ^2, is correctly specified in the sample size calculation. On the other hand, in trials with a time-to-event endpoint, this relationship depends on whether the true treatment effect exceeds the clinically meaningful effect and whether the hazard rates are correctly specified in the sample size calculation.

For a continuous endpoint, we have represented a clinically meaningful treatment effect as the difference in means, $\mu_1 - \mu_2$. Cohen[17] proposed expressing instead using the standardized effect size, $(\mu_1 - \mu_2)/\sigma$. Since the sample size formula (2) depends on $\mu_1 - \mu_2$ and σ^2 only through the standardized effect size, it is not necessary to specify the sample variance σ^2 separately when the clinically meaningful effect is defined in terms of the standardized effect size. However, a standardized effect size is meaningful and valid only when the specified variance is accurate. Furthermore, using a standardized effect size does not allow us to investigate the impact of a misspecified variance on the relationship between the probabilities of clinical significance and statistical significance.

Some investigators, such as Batterham and Hopkins[18] and Sullivan and Feinn,[19] argue that clinical significance is more important than statistical significance. However, this perspective overlooks the critical role of evidence in evaluating an observed treatment effect. Even if the observed effect is large, we cannot fully trust it if the individual treatment effects are highly variable and the sample size is small. This variability is why statistical significance is essential. Furthermore, the studies are typically designed to evaluate statistical significance rather than clinical significance. If clinical significance is considered important the study should be designed accordingly.

To avoid situations in which statistical significance is achieved despite clinical insignificance, it is important to ensure an appropriate, not excessively large, sample size. In observational studies, investigators often use very large sample sizes to include as many cases as possible. In such situations, the observed effect size should be carefully evaluated even if the statistical significance is high.

In general, however, statistical significance should be prioritized over clinical significance because an observed treatment effect can still be statistically unreliable if the standard error (SE) of the observed treatment effect is large. In prospective studies, it is more common to encounter statistical significance without clinical significance. Conversely, it is almost impossible to observe clinical significance without statistical significance unless the population variance of continuous endpoint is severely underestimated or the hazard rates of time to event endpoint are severely overestimated, leading to an overly underpowered sample size.

Conclusion

As pointed by Ranganathan et al.[5] and Willigenburg and Poolman,[20] a test for statistical significance is not designed to determine whether two patient groups differ by a clinically meaningful amount in an endpoint. Rather, it tests whether there is any difference in the endpoint between the groups. However, statistical significance and clinical meaningfulness become connected during sample size determination. Specifically, the sample size for a statistical test is calculated to ensure a high probability, called power, of detecting a pre-specified clinically meaningful difference between two groups.

If the sample size of a study is calculated using the true variance and treatment effect for continuous endpoints, or the true hazard rate of the control group and the true treatment effect for time-to-event endpoints, then it is possible to observe statistical significance without clinical significance, but not clinical significance without statistical significance.

The probability of achieving clinical significance is only 50% when the specified clinically meaningful treatment effect equals the true treatment effect, regardless of other factors such as sample size or variance. Consequently, in a reasonably designed study, the probability of clinical significance is generally smaller than that of statistical significance. An increased probability of clinical significance can occur only when the true treatment effect exceeds the specified clinically meaningful effect. In such cases, however, the probability of statistical significance will increase too.

Funding

This research received no external funding.

Author contributions

Conceptualization and supervision, S.H.J.; computation, C.J.; writing—original draft, S.H.J. and C.J. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declares that there are no conflict of interest.

Data availability statement

All data supporting the findings of this publication are available within this article.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Sample availability

No physical samples were used in this study.

Supplementary materials

There are no supplementary materials.

References

1. Armijo-Olivo S. The importance of determining the clinical significance of research results in physical therapy clinical research. *Braz J Phys Ther*. 2018;22:175–176.
2. Dahlberg SE, Korn EL, Le-Rademacher J, et al. Clinical versus statistical significance in studies of thoracic malignancies. *J Thorac Oncol*. 2020;15(9):1406–1408.
3. LeFort SM. The statistical versus clinical significance debate. *Image J Nurs Sch*. 1993;25:57–62.

4. Nahm FS. What the P values really tell us. *Korean J Pain*. 2017;30:241–242.

5. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: clinical versus statistical significance. *Perspect Clin Res*. 2015;6(3):169–170.

6. Sharma H. Statistical significance or clinical significance? A researcher's dilemma for appropriate interpretation of research results. *Saudi J Anaesth*. 2021;15(4):431–434.

7. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. *Int J Sports Phys Ther*. 2014;9(5):726–736.

8. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Austr Crit Care*. 2010;23(2):93–97.

9. Elasan S. The difference between clinical significance and statistical significance: an important distinction for clinical research. *Turk J Med Sci*. 2024;54(6):1419.

10. Machin D, Campbell MJ, Tan SB, et al. Sample Size for Clinical, Laboratory and Epidemiology Studies. 4th ed. Oxford, UK: Blackwell Science; 2018.

11. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc A*. 1972;135:185–206.

12. George SL, Desu MM. Planning the size and duration of a trial studying the time to some critical event. *J Chronic Dis*. 1974;27:15–24.

13. Schoenfeld DA. Sample size formula for the proportional hazards regression model. *Biometrics.* 1983;39:499–503.

14. Kwak M, Jung SH. Optimal two-stage logrank test for randomized phase II clinical trials. *J Biopharm Stat.* 2017;27(4):639–658.

15. Miller RG. Survival Analysis. New York, NY: John Wiley & Sons; 1981.

16. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J Chronic Dis*. 1981;34:469–479.

17. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

18. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50–57.

19. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ*. 2012;4:279–282.

20. Willigenburg NW, Poolman RW. The difference between statistical significance and clinical relevance: the case of minimal important change, non-inferiority trials, and smallest worthwhile effect. *Injury.* 2023;54:110764.