KJA
Korean Journal of Anesthesiology

Alternatives to the P value: connotations of significance

Junyong In, Dong Kyu Lee

Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, Goyang, Korea

Statistical Round

Korean J Anesthesiol 2024;77(3):316-325
https://doi.org/10.4097/kja.23630
pISSN 2005-6419 • eISSN 2005-7563

Received: August 20, 2023
Revised: November 26, 2023
Accepted: March 11, 2024

Corresponding author:
Dong Kyu Lee, M.D., Ph.D
Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, 27 Dongguk-ro, Ilsandong-gu, Goyang 10326, Korea
Tel: +82-31-961-7869
Fax: +82-31-961-7864
Email: entopic@dongguk.edu
ORCID: https://orcid.org/0000-0002-4068-2363

The statistical significance of a clinical trial analysis result is determined by a mathematical calculation and probability based on null hypothesis significance testing. However, statistical significance does not always align with meaningful clinical effects; thus, assigning clinical relevance to statistical significance is unreasonable. A statistical result incorporating a clinically meaningful difference is a better approach to present statistical significance. Thus, the minimal clinically important difference (MCID), which requires integrating minimum clinically relevant changes from the early stages of research design, has been introduced. As a follow-up to the previous statistical round article on P values, confidence intervals, and effect sizes, in this article, we present hands-on examples of MCID and various effect sizes and discuss the terms statistical significance and clinical relevance, including cautions regarding their use.

Keywords: Clinical relevance; Clinical significance; Confidence intervals; Effect size; Minimal clinically important difference; Patient outcome assessment; P value; Statistical significance; Statistics.

Introduction

When researchers review a paper, they expect to find scientific and clinically substantiated evidence for the effectiveness of the treatment of interest. To establish scientific evidence, employing a statistical approach encompassing sample size calculations is a common practice. The CONSORT statement includes statistical analysis-related content, such as sample size determination and appropriate statistical method selection, as well as research design components, such as randomization, blindness, and participant selection. Using the appropriate study design to acquire data becomes the foundation for statistical significance throughout statistical analysis. Because Pearson advocated for null hypothesis significance testing (NHST), a P value of 5% has become the threshold for determining statistical significance. Although the widely accepted significance level of 5% for the P value is an indicator of statistical significance, it does not constitute evidence of clinical relevance [1]. According to the International Committee of Medical Journal Editors (ICMJE) recommendations for medical journal standards [2],[1] clinical relevance refers to an effect of an intervention or treatment that promotes healing from a certain disease or has another similar positive influence, reduces the complication rate and illness duration, or consequently improves the quality of life. The value of these positive effects is appreciated from various points of view, and the clinical relevance is determined based on the results. We have already introduced confidence intervals and effect sizes in a previous statistical

[1] In the "Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals," published by the ICMJE in 2023, clinical relevance is expressed using the term clinical significance. In this article, we limit the use of the term significance to refer to statistical significance (a statistical term) and thus use the term clinical relevance rather than clinical significance.

round article entitled "Alternatives to P value: Confidence Interval and Effect size." When presented with a P value, these indices are good indicators of statistical significance and clinical relevance [3].

In this article, we introduce various types of effect sizes that can be used to describe statistical results and other indices that indicate clinical relevance, such as the minimal clinically important difference (MCID). In addition, we discuss how to interpret and describe statistical results using these indices, which encompass both clinical relevance and statistical significance.

Confidence intervals and effect sizes

Statistical significance is determined according to the decision criteria of NHST. NHST is a statistical method based on validating the null hypothesis (H0: the compared groups do not differ) to determine that "the comparative groups are not different," except with a type I error, which refers to the probability of accidentally observing a difference that is not there. This interpretation method does not indicate the direction or magnitude of the trial results; it simply indicates whether a statistical difference exists under the fixed significance level. In addition, the P value cannot be used to determine the magnitude of the difference because it is affected not only by the difference but also by the sample size and variability of the measured results [4]. A small absolute difference between central measures of groups can achieve statistical significance if the sample size is sufficiently large to create distinct or slightly overlapping distributions of observed data. Conversely, a large absolute difference between central measures of groups may not be statistically significant if the distributions substantially overlap due to a small sample size. Nonetheless, some researchers misinterpret the P value based on NHST as indicating something is "more significant" or "less significant." These incorrect interpretations have frequently been presented, even though researchers should know that comparing P values cannot be used to interpret the strength of significance because the NHST implies a dichotomous decision (whether the null hypothesis is true or false) [5]. Confidence intervals and effect sizes can be used to rectify such errors to express the magnitude of the differences or ratios observed in clinical trial results, as discussed in a previous article [2].

Confidence intervals

The confidence interval can be calculated to determine the range estimation using statistical probability. The representative value (e.g., mean), the measure of statistical dispersion (e.g., standard deviation), and the sample size of the group are used to estimate the confidence interval. Although the real mean of the target population is unknown, we can presume, with a preset probability, that the expected mean of a future population in the same environment with the same intervention will be located within this confidence interval. This statistical process enables us to expect an effect from an intervention in a future sample regardless of the unknown real value of the target population. Beyond determining the statistical significance based on whether the confidence interval includes a null value, using a confidence interval allows us to presume the potential direction and magnitude of the effect that may be observed in future patients from the same population who receive the same intervention. The confidence interval provides a range that reflects the statistical uncertainty of sampling and the statistical test process, which enables us to speculate on the expected results in real clinical situations. The P value represents the probability of accepting or rejecting the hypothesis, and the confidence interval represents the range of the estimated representative value along with the uncertainty (margin of error), where the real value of the population would exist [6]. However, the confidence interval is not a property of the observed data but rather a characteristic of a sampling distribution, such as the standard error of the mean (SEM). The sampling distribution is an imaginary distribution composed of the means of data that are repeatedly sampled from the population using the same method as that of the observed data. For example, the means from the groups are the observed values, and the confidence interval of the mean difference is a range estimated by probability and statistics based on the hypothesis. Similarly, the standard deviation, which indicates the dispersion of data, is the observed value, while the upper and lower limits of the confidence interval are statistically estimated values. The confidence interval cannot be interpreted as the mean and the standard deviation explaining the observed data distribution. The confidence interval is interpreted such that if the experiment is repeated using the same hypothesis and a confidence interval is calculated from each experiment, we can expect that the true population mean would fall within the given range of those intervals with a certain probability (usually 95%).

Compared to the dichotomous nature of the P value, including the confidence interval in the statistical result has the advantages described above. However, determining quantitative differences between clinical trials is frequently complex except in cases of mean difference or ratio comparisons.

Effect size

The effect size is a statistic representing the observed effect's standardized magnitude and direction. A detailed description of

Table 1. Statistical Methods and Recommended Effect Sizes

Statistical method	Effect size	Calculation	Effect size interpretation	
Student's t-test	Cohen's *d*	$d = \frac{\bar{X}_T - \bar{X}_C}{s_{pooled}}$ $\left(s_{pooled} = \sqrt{\frac{(n_T - 1){s_T}^2 + (n_C - 1){s_C}^2}{n_T + n_C - 2}}\right)$	$\bar{X}_T$, $\bar{X}_C$: Means of treatment and control groups S_{pooled}: Pooled SD n_T, n_C: Sample size of treatment and control groups S_T, S_C: SD of treatment and control groups	0.2 to < 0.5 Small effect 0.5 to < 0.8 Medium effect ≥ 0.8 Large effect
	Correlation coefficient r (BESD)	$r = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{d^2}{(d^2 + 4)}}$	*t*: Test statistics *df*: Degrees of freedom *d*: Cohen's *d*	0.1 to < 0.3 Small effect 0.3 to < 0.5 Medium effect ≥ 0.5 Large effect
Mann-Whitney rank sum test (Mann-Whitney *U* test)	*r*	$r = \frac{\lvert z \rvert}{\sqrt{n}}$	*z*: Test statistics *n*: Total sample size	0.1 to < 0.3 Small effect 0.3 to < 0.5 Medium effect ≥ 0.5 Large effect
	Vargha and Delaney's A	$VDA = \frac{U}{n_1 \times n_2}$	*VDA*: Vargha and Delaney's A *U*: U statistics n_1, n_2: Sample size of each group	0.56 to < 0.64, > 0.34 to 0.44 Small effect 0.64 to < 0.71, > 0.29 to 0.34 Medium effect ≥ 0.71, ≤ 0.29 Large effect
	Cliff's delta*	$\delta = 2(VDA - 0.5)$	δ: Cliff's delta *VDA*: Vargha and Delaney's A	0.11 to < 0.28 Small effect 0.28 to < 0.43 Medium effect ≥ 0.43 Large effect
Paired t-test	Cohen's *d*	$d = \frac{\bar{X}_{after} - \bar{X}_{before}}{S_{difference}}$	$\bar{X}_{after}$, $\bar{X}_{before}$: Mean of observations before and after $S_{difference}$: SD from differences in observations	0.2 to < 0.5 Small effect 0.5 to < 0.8 Medium effect ≥ 0.8 Large effect
Wilcoxon signed rank test (Wilcoxon Z test)	Matched-pairs rank biserial correlation coefficient	$r_c = \frac{4\lvert T - 0.5(R_+ + R_-)\rvert}{N(N+1)}$	r_c: Biserial correlation coefficient *T*: Smaller value between R_+ and R_- R_+, R_-: Sum of ranks with a positive or a negative sign *N*: Number of pairs	0.1 to < 0.3 Small effect 0.3 to < 0.5 Medium effect ≥ 0.5 Large effect
	r	$r = \frac{\lvert z \rvert}{\sqrt{n}}$	*z*: Test statistics *N*: Number of pairs	0.1 to < 0.3 Small effect 0.3 to < 0.5 Medium effect ≥ 0.5 Large effect
ANOVA	Coefficient η^2	$\eta^2 = \frac{SS_{ef}}{SS_t}$	SS_{ef}: Sum of squares for the effect† SS_t: Total sum of squares	0.02 to < 0.13 Small effect 0.13 to < 0.26 Medium effect ≥ 0.26 Large effect
	Partial η^2	$\eta^2_p = \frac{SS_{ef}}{SS_{ef} + SS_{er}}$	η^2_p: Partial eta-squared SS_{ef}: Sum of squares for the effect SS_{er}: Sum of squares error†	0.02 to < 0.13 Small effect 0.13 to < 0.26 Medium effect ≥ 0.26 Large effect

(Continued to the next page)

https://doi.org/10.4097/kja.23630

Table 1. Continued

Statistical method	Effect size	Calculation		Effect size interpretation	
	Coefficient ω^2 (for between-subject designs, Unbiased estimate of η^2)	$\omega^2 = \frac{df_{ef}(MS_{ef} - MS_{er})}{SS_t + MS_{er}}$	df_{ef}: Degrees of freedom for the effect MS_{ef}, MS_{er}: Mean square of the effect, error SS_t: Total sum of squares	0.01 to < 0.06 0.06 to < 0.14 ≥ 0.14	Small effect Medium effect Large effect
	Partial ω^2	$\omega^2_p = \frac{df_{ef}(MS_{ef} - MS_{er})}{df_{ef}MS_{ef} + (n - df_{ef})MS_{er}}$	ω^2_p: Partial ω^2 df_{ef}: Degrees of freedom for the effect MS_{ef}, MS_{er}: Mean square of the effect, error n: Sample size	0.01 to < 0.06 0.06 to < 0.14 ≥ 0.14	Small effect Medium effect Large effect
	Cohen's f	$f = \sqrt{\frac{\sum_{j=1}^{p}(\mu_j - \mu)^2/p}{\sigma^2}} = \sqrt{\frac{\eta^2}{1 - \eta^2}}$	p: Number of groups μ_j: Mean of each group μ: Mean of whole sample σ: Standard deviation of whole sample η^2: Coefficient η^2 (Cohen's f is interchangeable with η^2, as shown)	0.10 to < 0.25 0.25 to < 0.40 ≥ 0.40	Small effect Medium effect Large effect
Kruskal-Wallis ANOVA on ranks (Kruskal-Wallis H)	η^2	$\eta^2 = \frac{H - k + 1}{n - k}$	H: Kruskal-Wallis test statistic k: Number of groups n: Total number of observations	0.02 to < 0.13 0.13 to < 0.26 ≥ 0.26	Small effect Medium effect Large effect
	E^2 (Epsilon-squared)	$E^2 = \frac{H}{(n^2-1)/(n^2+1)}$	H: Kruskal-Wallis test statistic n: Total number of observations	0.01 to < 0.08 0.08 to < 0.26 ≥ 0.26	Small effect Medium effect Large effect
RM ANOVA	Partial η^2	$\eta^2_p = \frac{SS_{ef}}{SS_{ef} + SS_{er}}$	η^2_p: Partial eta-squared SS_{ef}: Sum of squares for the effect SS_{er}: Sum of squares error†	0.02 to < 0.13 0.13 to < 0.26 ≥ 0.26	Small effect Medium effect Large effect
	Generalized η^2	$\eta^2_G = \frac{SS_{ef}}{\delta \times SS_{sf} + \Sigma SS_{measured}}$	η^2_G: Generalized η^2 δ: 0 if the effect involves one or more measured factors; 1 if the effect involves only manipulated factors SS_{ef}: Sum of squares for the effect $SS_{measured}$: Sum of squares for the measurement‡	0.02 to < 0.13 0.13 to < 0.26 ≥ 0.26	Small effect Medium effect Large effect
	Partial ω^2	$\omega^2_p = \frac{df_{ef}(MS_{ef} - MS_{er})}{df_{ef}MS_{ef} + (n - df_{ef})MS_{er}}$	ω^2_p: Partial ω^2 df_{ef}: Degrees of freedom for the effect MS_{ef}, MS_{er}: Mean square of the effect, error n: Sample size	0.01 to < 0.06 0.06 to < 0.14 ≥ 0.14	Small effect Medium effect Large effect
Friedman RM ANOVA on ranks	Kendall's W (coefficient of concordance)	$W = \frac{\chi_w^2}{N(k-1)}$	χ_ω^2: Friedman test statistic N: Sample size k: Number of measurements per subject	0.1 to < 0.3 0.3 to < 0.5 ≥ 0.5 (see footnote§)	Small effect Medium effect Large effect

(Continued to the next page)

Table 1. Continued

Statistical method	Effect size	Calculation	Effect size interpretation	
Chi-square test	Cramér's V (Cramér's phi)¶	$\phi_c = \sqrt{\frac{\chi^2}{n(k-1)}}$	ϕ_c: Cramér's V (extended phi coefficient for 2 × 2 table)	k = 2
			χ^2: Chi-square statistic	0.1 to < 0.3 Small effect
			n: Sample size	0.3 to < 0.5 Medium effect
			k: Lesser value between numbers of column and row	≥ 0.5 (see footnote) Large effect
Fisher's exact test	Phi coefficient**	$\phi = \sqrt{\frac{\chi^2}{n}}$	χ^2: Chi-square statistic	0.1 to < 0.3 Small effect
			n: Sample size	0.3 to < 0.5 Medium effect
				≥ 0.5 Large effect
Two-proportions z-test	Cohen's h	$h = \lvert 2arcsin\sqrt{p_1} - 2arcsin\sqrt{p_2} \rvert$	p_1, p_2: Two given probabilities or proportions	0.2 to < 0.5 Small effect
			$2arcsin\sqrt{p}$: Arcsine transformation	0.5 to < 0.8 Medium effect
				≥ 0.8 Large effect
Correlation analysis	Pearson correlation coefficient r	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$	r: Correlation coefficient	0.1 to < 0.3 Small effect
			$\bar{x}$, $\bar{y}$: Means of x and y	0.3 to < 0.5 Medium effect
			x_i, y_i: Samples of variable x and y	≥ 0.5 Large effect
	Spearman's ρ	$r_s = \frac{cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$	r_s: Spearman' s ρ	0.1 to < 0.3 Small effect
			$cov(R(X), R(Y))$: Covariance of the rank variables	0.3 to < 0.5 Medium effect
			$\sigma_{R(X)}$, $\sigma_{R(Y)}$: Standard deviations of the rank variables	≥ 0.5 Large effect

The effect sizes listed in this table are not a complete list of available effect sizes. The authors chose some effect sizes according to their preference and recommendation. BESD: binomial effect size display [3], SD: standard deviation, ANOVA: analysis of variance, RM: repeated measures. *The original definition of Cliff's delta involves a pre-defined matrix and the following process [7]. Fortunately, Cliff's delta is linearly related to Vargha and Delaney's A [8]. Therefore, the interpretation of Cliff's delta is converted from the recommendation of Vargha and Delaney's A. †"effect" indicates the within- and between-subject factors. ‡Generalized η^2 is different from η^2 and partial η^2. Generalized η^2 is an effect size considering the interaction and has various formulas according to the study design. For details, refer to the article by Bakeman [9]. §Kendall's W uses Cohen's interpretation guidelines. Kendall's W is a test statistic of agreement between groups where W = 1 indicates that all groups have identical rank by the intervention, a complete agreement. That is, high Kendall's W represents concordant changes by the intervention (a repeated measures factor). ‖As mentioned above, Kendall's W is a statistic related to Friedman's ANOVA and represents the general effect of the overall ANOVA test. The effect size r from each multiple comparison (such as Bonferroni corrected Wilcoxon signed-rank tests) would be informative. ¶Guideline based on Cohen's suggestion. Alternatively, refer the Table 4 of the previously published article [3].

	Small effect	Medium effect	Large effect
k = 2	0.100 to < 0.300	0.300 to < 0.500	≥ 0.500
k = 3	0.071 to < 0.212	0.212 to < 0.354	≥ 0.354
k = 4	0.058 to < 0.173	0.173 to < 0.289	≥ 0.289
k = 5	0.050 to < 0.150	0.150 to < 0.250	≥ 0.250
k = 6	0.045 to < 0.134	0.134 to < 0.224	≥ 0.224

**Using the phi coefficient for the Fisher's exact test is controversial because it comes from the chi-square statistic. Using the odds ratio instead of the phi coefficient has been recommended. However, some articles still report using the phi coefficient or Cramér's V for the Fisher's exact test results. One problem with using the phi coefficient, Cramér's V, and Cohen's w is that they require uniformly distributed marginals for a 2 × 2 table.

https://doi.org/10.4097/kja.23630

the basic concept of effect size is provided in a previous article [3]. Table 1 summarizes the effect sizes corresponding to the different statistical analysis methods. Including the effect size in the statistical analysis results overcomes the limitations of the P value and enables descriptions of the quantitative and qualitative magnitude of the treatment effect, making it possible to compare the effects between groups or between trials and, thus, is the main statistic used in systematic reviews. The effect size is a point-estimated value, such as a mean, and a standardized value of an effect of a clinical trial intervention. In this respect, the advantage of the effect size is that it is more intuitive than the confidence interval and easy to interpret its meaning. Considering these advantages, an increasing number of studies have presented statistical results using effect sizes [10–13].

The effect size also has its own confidence interval based on the significance level, and there are often situations where interpretation of effect size using confidence intervals is necessary. Presenting the confidence intervals with the odds ratio (OR), relative risk, and area under the receiver operating characteristic curve is common practice. Various effect sizes can be calculated using R (R Core Team). The supplementary document presents the methods for calculating various effect sizes and the corresponding confidence intervals (Supplementary Materials 1 and 2).

Besides the effect size from a specific statistic, the number needed to treat (NNT; also, the number needed to treat for an additional beneficial outcome [NNTB]) is useful to describe the number of changes between comparison groups. The NNT describes the required number of patients to include in the treatment group to observe a beneficial effect in one patient from the treatment in a clinical trial. The NNT is an epidemiological measurement that is usually used to discuss the treatment effect of a certain medication. In clinical trials, the NNT is considered an index of effect, even though it is not a statistic like the effect size. A large NNT suggests that the experimental treatment is less effective because a large number of patients are required to obtain an effect from the treatment.

The NNT is the inversed value of the absolute risk reduction:

$$NNT = \frac{1}{(I_c - I_t)}$$

where I_c is the incidence of the control group and I_t is the incidence of the treatment group.

An example is as follows: a randomized controlled trial is conducted to investigate the preventative effect of Drug A on postoperative nausea and vomiting. The observed postoperative nausea and vomiting rates are 40% and 30% in the control and treatment groups, respectively. The absolute risk reduction is thus 10% (40%–30%) and the NNT is 10. This means that the preventive effect can be observed when 10 patients are treated with Drug A. Similar to the NNT, the number needed to harm (NNH) is an index of a hazardous effect and can also be useful for comparing effects.

Minimal clinically important difference (MCID)

In the early 20th century, MCID was introduced to measure clinical improvement in patients [14]. At the beginning of the 21st century, magnitude-based inference (MBI) was introduced in the field of sports medicine. MBI assesses the observed effects based on three criteria: harmful, trivial (small changes), and beneficial [15]. However, some scholarly journals no longer accept MBI as a valid statistical approach because it lacks a clear mathematical foundation and is associated with an increased risk of type I errors [16,17]. On the contrary, MCID is a statistical method of approaching the difference in the effects perceived by the patient in clinical settings rather than the numerical difference based on statistical significance. This measure is becoming increasingly common in statistical and medical research areas [18,19]. MCID is a representative method for determining clinical relevance that involves setting a specific value of the measured outcome as the threshold for meaningful effects.[2)] This threshold indicates the minimal amount for important or meaningful changes in the measured outcome to be observed in patients or participants, and changes in the outcome that are larger than this threshold are considered clinically relevant. However, no method is generally accepted as the standard for determining the threshold for clinically relevant changes. Several articles on determining MCID for outcomes in various situations have recently been published in various medical fields [20–24].

MCID is useful for assessing the clinical relevance of the outcome variable in participants. Particularly in pain research, patient-reported outcomes (PROs), such as the visual analog scale (VAS) or numerical rating scale (NRS), are commonly used, along with opioid consumption. The statistical analysis of these results contributes to the clinical values of the findings. However, statistically significant differences in these variables do not constitute an evaluation of the treatment effect as perceived by patients. MCID was thus introduced to define the treatment effect as perceived by the patients. In a study assessing the effect of pain management after surgery, for instance, different minimum thresholds for the clinical importance of pain relief may need to be set for patient

[2)] Besides MCID, several terms have been proposed, such as the minimal clinical difference (MCD), minimal clinically important improvement (MCII), and robust clinically important difference (RCID).

groups undergoing different types of surgery, such as abdominal or foot surgeries. Additionally, within the same study, the minimum threshold for judging the side effects of pain medications may vary depending on the severity of pain, as patients may tolerate side effects differently based on pain intensity. Therefore, interpreting differences in opioid consumption in the context of pain relief also differs based on clinical considerations and patient perceptions.

Interpreting MCID often involves testing the statistically significant proportion of patients who achieve a change equal to or greater than MCID in both the control and treatment groups using NHST. Alternatively, researchers may divide the study population into groups based on whether they exhibit a change equal to or greater than MCID and then statistically analyze the factors associated with the observed differences. This approach demonstrates how meaningful effects are observed in an actual patient population beyond simply presenting the differences in treatment effects between the two groups.

A standard method for calculating MCID has not been fully established. Some representative calculation methods include a distribution-based method based on the distribution of observed values and an anchor-based method that involves comparing generally accepted measurements as the standard (anchor) against an evaluation method that is widely used in clinics or more specific evaluation methods. As these methods have various limitations, a new method of comparing and coordinating the results of these methods to determine MCID, known as the triangulating method, has recently been attracting attention.[3] In addition, the Delphi method, which involves a panel of experts and patients reaching a consensus on the criteria through multiple rounds and determining MCID through a literature review, is another method. Unfortunately, none of these methods have been accepted as standard because each has advantages and disadvantages.

The distribution-based method follows a process similar to that used to determine the measurement error or effect size. For this method, factors such as the standard error of measurement ($SE_M = SD\sqrt{1-\text{Cronbach's }\alpha}$, it is different from the standard error of the mean, SEM), standard deviation, or the factors involved in these measurements are utilized. Commonly recommended distribution-based indicators include the SE_M, standard deviation of baseline (pretreatment) observations, 0.5 standard deviations, and the associated 1.96 SE_M which is related to the reliable change index (RCI). RCI can be calculated as the standard error of the difference in scores between two measurement methods, represented as $(x_1 - x_2)/\sqrt{2 \times SE_M^2}$, where x_1 and x_2 are scores from the respective measurement methods. If the calculated value is >1.96, it is considered significant at the 5% level, indicating that the change in measurement is "not likely due to a measurement error. Some researchers have also added additional criteria that they considered important regarding the characteristics of experimental design and data. Furthermore, based on the general interpretation of effect sizes, an effect size of 0.2, corresponding to a small effect, is sometimes set as MCID. With distribution-based methods, making a case for clinical significance is challenging because no clinical anchor is available to provide meaning to the criteria.

The anchor-based method involves using a reference assessment technique, such as the global assessment scale (GAS) or the global impression of changes (GIC), as an "anchor." For this method, values measured by individuals or researchers to estimate MCID are used (Fig. 1) [25]. This method involves comparing the mean change in response based on the assessment method under study (e.g., the VAS or NRS, which are PROs) with individual improvement effects based on a comprehensive assessment scale (anchor). This can also be determined using a receiver operating characteristic (ROC) curve analysis. This method begins with the assumption that the measurement method used is significantly associated with a reference evaluation method (anchor) chosen from among various assessment methods. Therefore, a relatively strong correlation is desirable (correlation coefficient $\geq$ 0.7). If a meaningful correlation is not observed, having strong confidence in the changes observed using the measurement values applied in a study is challenging. Furthermore, if a comprehensive assessment scale is evaluated retrospectively, relying on patients' memory may lead to recall bias. Therefore, re-validating the results using additional measurable criteria (anchors), such as analgesic consumption, to mitigate potential bias is advised [26].

Given that multiple methods are available to determine MCID, a variety of MCID values can be calculated for the same patients in the same situation. Although comparing and reconciling the outcomes of various methods using the triangulating method is advised, a comprehensive systematic approach for this purpose has not yet been established. The process typically begins with a distribution-based analysis, followed by a supplementary evaluation using a comprehensive assessment scale. Subsequently, among MCID values identified using these two methods, researchers often employ ROC analyses to determine the most appropriate choice, or they may opt for the average of these values as the final MCID [27,28]. Given the lack of established statistical methods for these procedures, the recommended approach for establishing the robustness of the selected MCID value involves

[3]The triangulating method is a measurement technique that utilizes the properties of triangles for measurement, known as triangulation. Here, it is used to determine a more accurate and reliable MCID through deciding between two different MCID values.

https://doi.org/10.4097/kja.23630

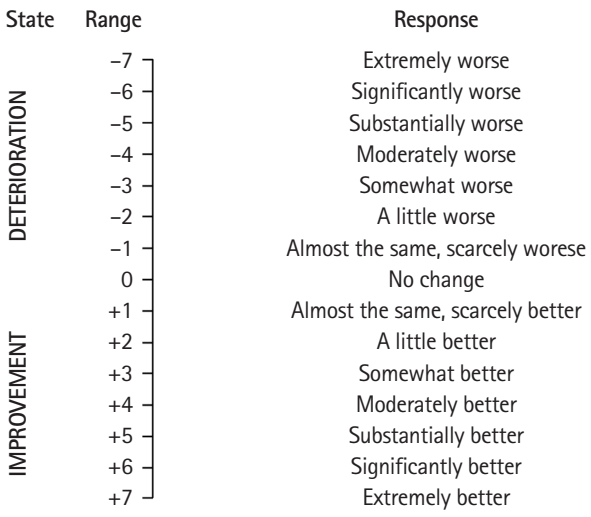

Fig. 1. An example of the Global Assessment Scale (GAS). The term 'moderately' suggests that changes are noticeable but not dramatic. 'Substantially' indicates that the changes are considerable and impactful. 'Significantly' denotes that the changes have far exceeded what was anticipated.

conducting a sensitivity analysis. This analysis assesses the impact of assumptions on outcomes, examines how changes in these assumptions affect results, and identifies uncertainties in research design and data collection processes. By evaluating their impact on final outcomes, researchers can verify the consistency of research findings under different conditions and enhance the reliability of their results. In the supplementary document, the overall process for calculating MCID along with an example of conducting the sensitivity analysis is presented.

Additional considerations must also be taken into account. First, the fundamental notion of "minimal" must be adequately considered. Determining whether the measurement tool used (e.g., the VAS or NRS) adequately captures the minimum amount of change is essential. If the instrument does not reflect the minimum change reported by the patient, this can lead to increased errors due to inaccurate measurements. Furthermore, the minimum change reported by patients is influenced by their perception thresholds; thus, measurement tools that can capture this should be employed. Research in the field of psychophysiology indicates that the minimum change based on patient reports is approximately 0.5 standard deviations (SD) of the effect size. This value is often used in anchor-based methods that rely on reference-assessment techniques. Second, even for results obtained using the same measurement tool, various factors influence MCID value, such as the study setting, participants, and the method of calculation. Therefore, applying an established MCID from one specific study to a clinical trial is challenging. To use a previously reported MCID, researchers must assess and consider the differences between the circumstances of the research conducted to determine MCID and the current one being conducted.

Clinical relevance vs. statistical significance

As discussed previously, the presence of statistical significance in the data analysis does not imply clinical relevance. Conversely, the absence of statistical significance does not necessarily mean a lack of clinical relevance. The latter scenario often arises when a study is conducted with an inadequate sample size or when the method of measuring the outcome variable exhibits significant variability. For instance, if the severity of postoperative nausea is recorded by the patients themselves using an NRS, patients' subjectivity cannot be entirely eliminated. Depending on the circumstances, similar levels of nausea symptoms could be measured differently. Consequently, even if an antiemetic with a potentially meaningful effect is administered, statistical significance may not be attainable. The concept of clinical relevance lacks an agreed-upon definition, and traditionally, many studies have assessed clinical relevance based on statistical significance.

The clinical effects reflected by the effect size estimate the average effect observed in the experimental group due to the intervention, allowing for the interpretation of the magnitude of the effect. However, the effect size does not provide a specific indication of how much an individual can expect to benefit; rather, it captures the overall effect, encompassing the entire group. Furthermore, the effect size is a dimensionless comparative measure, which means interpreting it directly at an individual level is challenging. One benefit of MCID is that it employs the same units of measurement as the actual variable, enabling assessment of clinical relevance for specific patients. It can serve as a reference for deciding whether to continue the current treatment or consider alternative approaches in individual patients within a clinical context. Essentially, it enables assessment at the individual level. Furthermore, integrating MCID in research facilitates its application in the evaluation of novel treatment methods.

However, MCID also has several limitations. In addition to the issues of bias and lack of established calculation methods mentioned earlier, the assessment of treatment effects through MCID can vary depending on the patient's circumstances or past experiences. For example, required MCID may be higher if a patient experienced higher pain levels before treatment. Similarly, if a patient has repeatedly encountered similar types of pain in previous experiences, a greater effect might be required to achieve a state of comfort. Additionally, because MCID represents the smallest

meaningful effect, it may not be suitable as a criterion for judging meaningful outcomes of a particular clinical intervention that aims to achieve substantial treatment efficacy [29]. The limitations of MCID continue to be evident. Anchor-based methods are often difficult to conduct given the lack of appropriate anchor measurements for a wide range of cases. Distribution-based methods frequently result in findings that lack clinical meaning for the chosen criteria and variable MCID values owing to changing criteria with every sample extraction, even when the same study is repeated.

Conclusion

Researchers and medical practitioners have developed new treatments and medications based on accumulated evidence from clinical trials, aiming to offer patients the best possible care. Research outcomes that demonstrate statistically significant differences are the strongest evidence provided that they are sufficiently clinically relevant. The statistical significance of a research outcome is determined through a binary decision-making process that involves a mathematic calculation based on the null hypothesis, which states that no difference or effect is present. However, clinical decision making requires constructive information on the expected effect of the treatment or medication, beyond the presence or absence of an effect. The effect size is a standardized statistic (value) of the magnitude and direction of change observed in a study, and MCID is a robust threshold for determining clinical relevance.

By combining metrics of clinical relevance, such as the effect size and MCID, with the conventional application of statistical significance and presenting outcomes derived from robust research designs, it becomes imperative that we can establish a foundation that holds both scientific and clinical significance. This approach has the potential to enhance our understanding not only from a scientific standpoint but also in terms of the practical clinical implications.

Funding

None.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Data Availability

The datasets generated during and analyzed during the current article are available as supplementary material 2.

Author Contributions

Junyong In (Writing – original draft; Writing – review & editing)
Dong Kyu Lee (Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Writing – original draft; Writing – review & editing)

ORCID

Junyong In, https://orcid.org/0000-0001-7403-4287
Dong Kyu Lee, https://orcid.org/0000-0002-4068-2363

Supplementary Materials

Supplementary Material 1. Examples of the effect size and MCID calculation.
Supplementary Material 2. Sample data.

References

1. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016; 70: 129-33.
2. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals 2023 [Internet]. International Committee of Medical Journal Editors. Available from http://www.icmje.org/icmje-recommendations.pdf.
3. Lee DK. Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol 2016; 69: 555-62.
4. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31: 337-50.
5. Kwak S. Are Only p-values less than 0.05 significant? A p-value greater than 0.05 is also significant! J Lipid Atheroscler 2023; 12: 89-95.
6. Benjamini Y, Veaux RD, Efron B, Evans S, Glickman M, Graubard BI. The ASA president's task force statement on statistical significance and replicability. Ann Appl Stat 2021; 15: 1084-5.
7. Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull 1993; 114: 494-509.
8. Vargha A, Delaney HD. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. J

https://doi.org/10.4097/kja.23630

Educ Behav Stat 2000; 25: 101-32.
9. Bakeman R. Recommended effect size statistics for repeated measures designs. Behav Res Methods 2005; 37: 379-84.
10. Libster R, Pérez Marc G, Wappner D, Coviello S, Bianchi A, Braem V, et al. Early high-titer plasma therapy to prevent severe Covid-19 in older adults. N Engl J Med 2021; 384: 610-8.
11. Concannon P, Rich SS, Nepom GT. Genetics of type 1A diabetes. N Engl J Med 2009; 360: 1646-54.
12. Hays J, Ockene JK, Brunner RL, Kotchen JM, Manson JE, Patterson RE, et al. Effects of estrogen plus progestin on health-related quality of life. N Engl J Med 2003; 348: 1839-54.
13. Chow JT, Turkstra TP, Yim E, Jones PM. The degree of adherence to CONSORT reporting guidelines for the abstracts of randomised clinical trials published in anaesthesia journals: a cross-sectional study of reporting adherence in 2010 and 2016. Eur J Anaesthesiol 2018: 942-8.
14. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407-15.
15. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. Int J Sports Physiol Perform 2006; 1: 50-7.
16. Sainani KL. The problem with "Magnitude-based Inference". Med Sci Sports Exerc 2018; 50: 2166-76.
17. Sainani KL, Lohse KR, Jones PR, Vickers A. Magnitude-based Inference is not Bayesian and is not a valid method of inference. Scand J Med Sci Sports 2019; 29: 1428-36.
18. Lemieux J, Beaton DE, Hogg-Johnson S, Bordeleau LJ, Goodwin PJ. Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. J Clin Epidemiol 2007; 60: 448-55.
19. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharm Stat 2004; 14: 97-110.
20. Todd KH, Funk KG, Funk JP, Bonacci R. Clinical significance of reported changes in pain severity. Ann Emerg Med 1996; 27: 485-9.
21. Danoff JR, Goel R, Sutton R, Maltenfort MG, Austin MS. How much pain is significant? Defining the minimal clinically important difference for the visual analog scale for pain after total joint arthroplasty. J Arthroplasty 2018; 33: S71-5.
22. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005; 64: 29-33.
23. Howard R, Phillips P, Johnson T, O'Brien J, Sheehan B, Lindesay J, et al. Determining the minimum clinically important differences for outcomes in the DOMINO trial. Int J Geriatr Psychiatry 2011; 26: 812-7.
24. Powell CV, Kelly AM, Williams A. Determining the minimum clinically significant difference in visual analog pain score for children. Ann Emerg Med 2001; 37: 28-31.
25. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994; 47: 81-7.
26. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. Pain 2000; 88: 287-94.
27. Myles PS, Myles DB, Galagher W, Boyd D, Chew C, MacDonald N, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. Br J Anaesth 2017; 118: 424-9.
28. Malec JF, Ketchum JM. A standard method for determining the minimal clinically important difference for rehabilitation measures. Arch Phys Med Rehabil 2020; 101: 1090-4.
29. Muñoz-Leyva F, El-Boghdadly K, Chan V. Is the minimal clinically important difference (MCID) in acute pain a good measure of analgesic efficacy in regional anesthesia? Reg Anesth Pain Med 2020; 45: 1000-5.