Revista Odonto Ciência
Journal of Dental Science

LITERATURE REVIEW

Rev Odonto Cienc 2018;33(1):84-90
http://revistaseletronicas.pucrs.br/ojs/index.php/fo
http://dx.doi.org/10.15448/1980-6523.2018.1.29437

Open Access

Effect size: a statistical basis for clinical practice

Leandro Almeida Nascimento Barros[a], Carolina Ferrari-Piloni[a], Érica Miranda Torres[b], Carlos Estrela[c], José Valladares-Neto[d]

ABSTRACT

OBJECTIVE: Effect size (*ES*) is the statistical measure which quantifies the strength of a phenomenon and is commonly applied to observational and interventional studies. The aim of this review was to describe the conceptual basis of this measure, including its application, calculation and interpretation.

RESULTS: As well as being used to detect the magnitude of the difference between groups, to verify the strength of association between predictor and outcome variables, to calculate sample size and power, *ES* is also used in meta-analysis. *ES* formulas can be divided into these categories: I – Difference between groups, II – Strength of association, III – Risk estimation, and IV – Multivariate data. The *d* value was originally considered small ($0.20 > d \leq 0.49$), medium ($0.50 > d \leq 0.79$) or large ($d \geq 0.80$); however, these cut-off limits are not consensual and could be contextualized according to a specific field of knowledge. In general, a larger score implies that a larger difference was detected.

CONCLUSION: The *ES* report, in conjunction with the confidence interval and *P* value, aims to strengthen interpretation and prevent the misinterpretation of data, and thus leads to clinical decisions being based on scientific evidence studies.

Keywords: effect size; *P* value; statistical interpretation; clinical decision-making; clinical effectiveness.

[a] Graduate student, School of Dentistry, Federal University of Goiás, Goiás, Brazil
[b] Associated Professor, Division of Prothesis and Scientific Methodology; School of Dentistry, Federal University of Goiás
[c] Head, Division of Endodontics, School of Dentistry, Federal University of Goiás, Goiás, Brazil
[d] Associated Professor, Division of Orthodontics, School of Dentistry, Federal University of Goiás, Goiás, Brazil

Tamanho do efeito: base estatística para a prática clínica

RESUMO

OBJETIVO: O tamanho do efeito (ES) é a medida estatística que quantifica a força de um fenômeno e é comumente aplicada a estudos observacionais e de intervenção. O objetivo desta revisão foi descrever a base conceitual desta medida, incluindo sua aplicação, cálculo e interpretação.

RESULTADOS: Além de ser usado para detectar a magnitude da diferença entre os grupos, para verificar a força da associação entre variáveis preditoras e de desfecho, para calcular o tamanho da amostra e a potência, ES também é usado em metanálise. As fórmulas ES podem ser divididas nestas categorias: I – Diferença entre grupos, II – Força de associação, III – Estimativa de risco e IV – Dados multivariados. O valor *d* foi originalmente considerado pequeno ($0,20 > d \leq 0,49$), médio ($0,50 > d \leq 0,79$) ou grande ($d \geq 0,80$); entretanto, esses limites de corte não são consensuais e podem ser contextualizados de acordo com um campo específico de conhecimento. Em geral, uma pontuação maior implica que uma diferença maior foi detectada.

CONCLUSÃO: O ES, em conjunto com o intervalo de confiança e valor de *P*, visa reforçar a interpretação e evitar a má interpretação dos dados, e, assim, leva a decisões clínicas baseadas em estudos de evidências científicas.

Palavras-chave: tamanho do efeito; valor de *P*; interpretação estatística; tomada de decisão; eficácia clínica.

Correspondence:
José Valladares-Neto
jvalladares@uol.com.br

Received: December 13, 2017
Accepted: January 14, 2019

Conflict of Interests: The authors state that there are no financial and personal conflicts of interest that could have inappropriately influenced their work.

INTRODUCTION

Effect size (*ES*) is the statistical measure which quantifies the strength of a phenomenon [1]. It is also known as effect magnitude and applies to different epidemiological designs, including observational and interventional studies. This estimate measures the magnitude of the difference between groups, the strength of the association between variables, and the risk of occurrence for a given event. As this measure is standardized, it allows one to compare estimates between different studies and is used in a pooled manner in meta-analyzes [2].

Hypothesis tests and the concept of *P* value were the main statistical tools used to analyze the strength of scientific evidence in quantitative studies throughout the twentieth century. However, the use of the *P* value as a statistical reference to show the effectiveness of treatments has been questioned in several publications [3, 4, 5], culminating in the 2016 American Statistical Association's recommendation of avoiding conclusions based exclusively on *P* values [6]. The classical methodology presents certain adverse characteristics, such as: 1) low reproducibility; 2) dependence on sample size and variance; 3) frequent clinical inconsistency; 4) use of arbitrary cut-off values ($P>0.05$ and $P<0.05$ as a result, to accept or reject the null hypothesis, respectively); and, 5) presenting dichotomous results only, that is, statistically significant or non-significant [7, 8, 9, 10]. A more appropriate interpretation includes knowing how much one intervention or association is better or greater when compared to another, and not simply whether or not there is a difference or association [11]. The concept of *ES* serves to fill this gap. Statistical significance and *ES* are currently complementary, and it is recommended that they be applied together, especially for the analysis and interpretation of primary outcomes [8, 12, 13].

Despite the fact that the *ES* estimate is related to relevant information and that its description has been widely recommended when reporting *P* value, few studies have explored or applied this concept to the health field [10]. In that light, the purpose of this article is to describe the conceptual basis of the most common measures of *ES*, and present information on its application, calculation and interpretation, with a view to helping understand the conclusion of studies and thus lead to improved clinical practice.

ES APPLICATION

Data interpretation

ES can be applied to different study designs. In clinical trials, it is used to detect whether one intervention is better or worse when compared to another, and not simply whether or not there is a difference, as explored by the concept of statistical significance [11]. It thus assists in making a clinical decision about the superiority (or otherwise) of a given intervention. When the *ES* is large enough, it differentiates between two treatments or decides whether one is preferable to another, from a clinical point of view [12]. In observational studies, *ES* can be understood as the strength of association between outcome and predictor variables and suggests not only whether there is an association but also how magnified it is [7, 14]. In addition, *ES* can also be expressed by relative risk [7, 15].

Meta-analysis

In meta-analysis, *ES* is extracted from individual primary studies, either directly or from data transformation, and then pooled to synthesize a standardized measure with greater statistical power [10, 16]. In this way, the greater accuracy provided by the joint data can be used to resolve controversies between primary studies and give an objective estimate of the scientific evidence. However, the interpretation and comparison of *ES* require careful consideration of the sources of variability [13].

Sample size calculation and statistical power

ES estimation is applied for calculating sample size (*n*) and statistical power (1-β) [17]. For the purpose of calculating a reasonable sample size, *ES* can be estimated by similar article published by others, pilot study results, or the minimum difference that would be considered important by experts [2]. Improper *n* affects the veracity of *P* values, and compromises internal validity. An undersized sample increases the probability of a type II error (β), while an oversized sample (big data) increases the probability of a type I error. Thus, large samples can give rise to a reduced *P* value. Thereby exaggerating the importance of the difference between interventions or associations [18]. For this reason, *n* calculation should be performed *a priori*. Otherwise, a *post hoc* statistical power analysis could confirm, or not, the validity of the study [15].

Statistical power is the probability of correctly rejecting the null hypothesis [4]. It can be influenced by three factors: level of significance (α), sample size, and *ES* [10]. As a general rule, the smaller the variance and the greater the *ES* and power, the smaller the sample size, and vice versa [19]. In other words, a strong association between two variables or a large existing difference will be easily detected in the sample, so that a small sample will be able to demonstrate this effect. To be detected, however, a weak association or small difference will require a larger sample size and power [20].

Calculating *ES*

ES estimates can be calculated by various formulas linked to different statistical tests, and it is advisable to report which formula was used when mentioning *ES* scores [11, 13, 10]. Most commonly used measures have been grouped according to the following categories: I – Group difference, II – Association strength, III – Risk estimation, and IV – Multivariate data (**Table 1**).

I – Group difference

This category evaluates the difference between means or frequencies when two or more groups are involved.

Table 1. Formulas for effect size estimative

Formulas	Variations
d of Cohen: $d = \frac{M_1 - M_2}{\frac{(n_1-1)DP_1^2+(n_2-1)DP_2^2}{n_1+n_2-2}}$	Combined standard deviation M_1 = mean of the experimental group DP_1 = standard deviation of the experimental group n_1 = sample size of the experimental group M_2 = mean of the control group DP_2 = standard deviation of the control group n_2 = sample size of the control group
$\sigma = \frac{\mu_1 - \mu_2}{\sigma}$	Population standard deviation μ_1 = population mean of the experimental group μ_2 = population mean of the control group σ = population standard deviation
d_m $d_m = \frac{M_1 - M_2}{\frac{DP_1+DP_2}{2}}$	Mean of the standard deviation M_1 = mean of the experimental group M_2 = mean of the control group DP_1 = standard deviation of the experimental group DP_2 = standard deviation of the control group
g of Hedges: $g = \frac{M_1 - M_2}{\frac{(n_1-1)DP_1^2+(n_2-1)DP_2^2}{n_1+n_2-2}} \cdot \left(1 - \frac{3}{4gl-1}\right)$	Combined standard deviation M_1 = mean of the experimental group DP_1 = standard deviation of the experimental group n_1 = sample size of the experimental group M_2 = mean of the control group DP_2 = standard deviation of the control group n_2 = sample size of the control group gl = degree of freedom (n-1)
Δ of Glass: $\Delta = \frac{M_1 - M_2}{DP_{control}}$	M_1 = mean of the experimental group M_2 = mean of the control group $DP_{control}$ = standard deviation of the control group
$\eta^2 = \frac{SS_E}{SS_T}$	SS_E = sum of squares for the exposure variable SS_T = total variance of outcome variables
${}_{partial}\eta^2 = \frac{SS_E}{SS_E + SS_{ER}}$	SS_E = sum of squares for the exposure variable SS_R = sum of squared errors
$\varepsilon^2 = \frac{SS_B - df_b MS_R}{SS_T}$	SS_B = sum between group effect SS_T = total variance MS_R = mean square of residuals df_B = degree of freedom between groups
$\omega^2 = \frac{SS_B - (df_b)\, MS_R}{SS_T + MS_R}$	SS_B = sum between group effect SS_T = total variance MS_R = mean square of residuals df_B = degree of freedom between groups
$phi\ (\phi) = \sqrt{\frac{x^2}{n}}$	x^2 = qui-square of independence n = sample size
V of Cramér: $V = \sqrt{\frac{x^2}{n(df_s)}}$	x^2 = qui-square of independence n = sample size df = lower degrees of freedom for number of rows and columns
r_{xy} of Pearson: $r_{xy} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$	$\sum$ = sum x and y = dependent variable value $\bar{x}$ and $\bar{y}$ = simple arithmetic means of the x and y values
r of Spearman: $r_s = 1 - \left(\frac{6\sum_{i=1}^{n} d_i^2}{N^3 - N}\right)$	R_i and S_i = sorted by ranks $\bar{R}$ and $\bar{S}$ = R and S mean (average) n = number of pairs $\sum$ = sum
$R^2 = 1 - \frac{SS_E}{SS_T}$	SS_E = squares sum for the effects SS_T = data variance
$f^2 = \frac{R_{AB}^2 + R_A^2}{1 - R_{AB}^2}$	R_B = interest variable R_A = other variables R_{AB} = combined ratio
$OR = \frac{(X_1Y_1)(X_0Y_0)}{(X_0Y_1)(X_1Y_0)}$	X = outcome probability in the treatment group Y = outcome probability in the control group
$RR = \frac{(X_1Y_1)\,/\,(X_1Y_1) + (X_0Y_0)}{(X_0Y_1)\,/\,(X_0Y_1) + (X_1y_0)}$	Exposure (x) and outcome (y) association
$\eta^2_{ajusted} = \frac{SS_E}{SS_E + SS_R}$	SS_E = squares sum for the effects SS_R = residual sum
$r^2_{ajusted} = 1 - \left[\frac{(1-R^2)-(n-1)}{n-k-1}\right]$	n = sample size k = number of predictors variables R^2 = squared correlation coefficient

The difference between means can be expressed in absolute or standardized terms. The simple difference between two means, for example, is the absolute difference, while the standardized is dimensioned by the variability, and so is more suitable for the comparison of multiple studies [2].

I.1 Difference of mean between two groups

The *d, g* and *Δ* formulas are used for this purpose. They have similar numerators (difference between absolute means) but different denominators, and are the population, combined, and control group standard deviations [10].

d

This measure was proposed by Cohen in 1962 and represents the most commonly used standardized mean difference [1]. It requires the following assumptions: normal distribution, unpaired groups, and variables measured on a continuous scale [1]. This index is used when the population standard deviation is known (δ formula) or when the standard deviation is used in a pooled manner (*d* formula) (**Table 1**) [10]. The *d* mean (dm) index is applied for the paired *t* test, using the means of the standard deviations between the groups [7].

g

This measure was proposed by Hedges in 1982 [10, 21]. The formula uses the pooled sample standard deviation and adds the correction factor J=(1-3/(4gl-1). This approach is commonly used for groups with different sample sizes, and is also used in small sample cases ($n<20$) or when the population standard deviation is unknown. It is commonly used for the *t*-test and meta-analysis.

Delta (Δ)

This measure was proposed by Glass in 1976 [10, 22]. It is an alternative approach to both *d* and *g*, when the standard deviations of the groups are significantly different. In this case, the standard deviation from the control group is chosen rather than a combined standard deviation.

I.2 Difference of mean between more than two groups

Eta squared (η^2) and partial eta squared ($_{partial}\eta^2$)

η^2 was introduced by Fischer in 1925 [8]. η^2 and $_{partial}\eta^2$ are the most commonly reported ES estimations for one-way analysis of variance (1-way ANOVA). They were drafted to compare three or more groups measured by continuous variables [19]. η^2 becomes biased as sample size decreases, caused by the lack of a population correction factor [11]. Also, with multiple factors tends to underestimate ES as the number of factors increases. For this reason, the partial eta will be better indicated [11,23]. It should be emphasized that the preference of use is $\eta^2 > {}_{partial}\eta^2 > \omega^2$ accordingly to the sample size [23].

Epsilon squared (ε^2) and omega squared (ω^2)

ε^2 and ω^2 were proposed by Kelley in 1935 and Hays in 1963, respectively [8, 11]. These are estimates of the effect size provided by the ANOVA test which use population correction factor. ω^2 and ε^2 are more conservative compared to η^2, thereby reducing the bias of small samples [11]. ω^2 is not appropriate for comparing groups with reduced samples. In such cases, the η^2 should be used [24]. Both ω^2 and ε^2 are better suited to compare ES between studies with the same experimental design [25].

I.3 Frequency difference

This subcategory is used for nominal qualitative data. It compares difference of frequency between groups, and the typical statistical test involved is the chi-square test.

phi (ϕ)

ϕ correlation coefficient was proposed by Karl Pearson and is used in a 2×2 contingency table. The values range from -1 to 1 [13].

V

In 1946, this measure was proposed by Harald Cramér, who extended the ϕ to larger contingency tables (Table 3×2, 2×4, 5×3...) [26]. The values range from 0 to 1, and the greater association between the variables is closer to 1.

II – Association strength

This category evaluates the strength of the shared variance between two variables (predictor and outcome).

Pearson (*r*) and Spearman (r_s) coefficient correlation

r was proposed by Karl Pearson and r_s by Charles Spearman. *r* measures the association strength between two continuous variables, and the usual normality and homoscedasticity assumptions are assumed. r_s is the non-parametric version, indicated for ordinal and continuous non-normal distribution variables. Neither have any unit of measurement, and range from -1 to +1 [26, 27].

R^2

R^2 is called the "coefficient of determination", also referred as r^2 or *r*-squared. It is calculated as the square of the *r*, ranges from 0 to 1, and is used in regression analysis [7]. It is the square of the *r* and ranges from 0 to 1. R^2 provides the value as a percentage when multiplied by 100 meaning the percentage of the variance of either variable is shared with the other variable.

f^2 (Cohen)

f^2 is recommended for comparing more than two groups by means of repeated measurements in regression and linear hierarchy models, whose objective is to evaluate the relationship of the variable with the outcome. This is represented by a global effect model where *R* is the correlation between the dependent and independent variables (both continuous) [28]. The local effect size can be estimated by modifying the global formula, where RB represents the variable of interest and RA the other variables. RAB is the combined ratio [28].

III – Risk estimation

This category compares the chance or risk for an outcome between two or more groups [2]. The score 1 represents no effect.

Odds Ratio (OR)

OR is an association measure applied to transverse and retrospective (case control) designs. It is reported as an ES index obtained from contingency tables (association between exposure and outcome) [29]. *OR* represents the ratio of chance of occurrence against the chance of non-occurrence of a determining event [23].

Relative Risk (RR)

RR is an association measure applied to prospective designs (clinical trials and cohort studies). It comprises a ratio of incidence observed in exposed and non-exposed groups [7]. The score ranges from 0 to 1 and can be transformed into a percentage when multiplied by 100.

IV – Multivariate data

This category deals with multivariate analysis, which plays a crucial role in understanding complex data sets which require a simultaneous examination of various variables.

Adjusted Eta squared ($n^2_{adjusted}$)

$n^2_{adjusted}$ is used for comparing three or more groups with predictor (independent) variables in multivariate analysis (MANOVA and ANCOVA). Its main advantage in relation to η^2 is in analyzing the effect of a specific variable while controlling the effect of other variables in the study (Hays, 1994). It has also been proposed for improving the comparability of *ES* findings between studies with the same methodological design.

Adjusted *R* squared ($R^2_{adjusted}$)

R^2 has a corrected variation called $R^2_{adjusted}$, which seeks to correct the variance errors shared by multiple predictors, used in multiple regression [11].

INTERPRETING ES

The extraction of maximum useful information from statistical research data helps the researcher to interpret results. *ES* estimates describe the observed effect and approaches to the practical relevance of the study. In addition, in terms of the statistical significance test, they emphasize the power of the tests and reduce the random error of a mere sample variation. In general, the larger the size, the larger the effect and impact caused by the variable under study. It should be noted that the effect size being sought is that of the population, but as this value is not available the sample effect size should be used to estimate the probable effect size of the population [15].

There is no consensus as to what constitutes a small, moderate, or large *ES*, because there are many different means of calculation and variations differ depending on the field of investigation [1,29]. According to Cohen (1988), pain values were small ($0.20 > d \leq 0.49$), medium ($0.50 > d \leq 0.79$) and large ($d \geq 0.80$) (**Table 2**) (**Figure 1**) [1].

Table 2. Common effect size uses and values

Category	Related statistical test	Effect size statistic	Values*		
			Small	Medium	Large
Difference between groups					
– Difference of 2 means	*t* test (similar SD)	Cohen´s *d*	0.20	0.50	0.80
	t test (known populational SD)	Cohen´s δ	0.20	0.50	0.80
	paired *t* test	Cohen´s *dm*	0.20	0.50	0.80
	t test (groups of different sizes)	Hedge´s *g*	0.20	0.50	0.80
	t test (different SD and heteroscedastic)	Glass´s Δ	0.20	0.50	0.80
– Difference of >2 means	ANOVA	η^2 (squared eta)	0.01	0.06	0.14
	ANOVA (groups of same size)	ω^2 (squared omega)	0.01	0.06	0.14
	ANOVA	ε^2 (squared epsilon)	0.01	0.06	0.14
	Kruskal-Wallis and Friedman tests	Cohen´s *f*	0.10	0.25	0.40
– Difference of frequencies	Chi-square (χ^2), 2 × 2 contingency table	ϕ (phi)	0.10	0.30	0.50
	Chi-square (χ^2), larger contingency table	Cramér´s *V*	0.10	0.30	0.50
Strength of association	Correlation (parametric)	Pearson´s *r*	0.10	0.30	0.50
	Correlation (non-parametric)	Spearman´s r_s / Kendal´s *t*	0.10	0.30	0.50
	Coefficient of determination	R^2	0.04	0.25	0.65
	Simple regression	Cohen´s *f*	0.10	0.25	0.40
	Logistic regression	*OR* / *RR*	1.50	2.00	3.00
	Poisson regression	*OR* / *RR*	1.50	2.00	3.00
Risk estimates	Odds Ratio (*OR*)	*OR*	1.50	2.00	3.00
	Relative Risk (*RR*)	*RR*	1.50	2.00	3.00
Multivariate data	MANOVA, MANCOVA	Cohen´s *f*	0.10	0.25	0.40
	MANOVA, MANCOVA	*Ajusted* η^2	0.01	0.06	0.14
	Multiple regression	*Ajusted* R^2	0.01	0.06	0.14

* Variable according to the decision context and comparative value of specific research area.
SD (standard deviation); ANOVA (analysis of variance); ANCOVA (Covariance Analysis); MANOVA (Multivariate Analysis of Covariance).

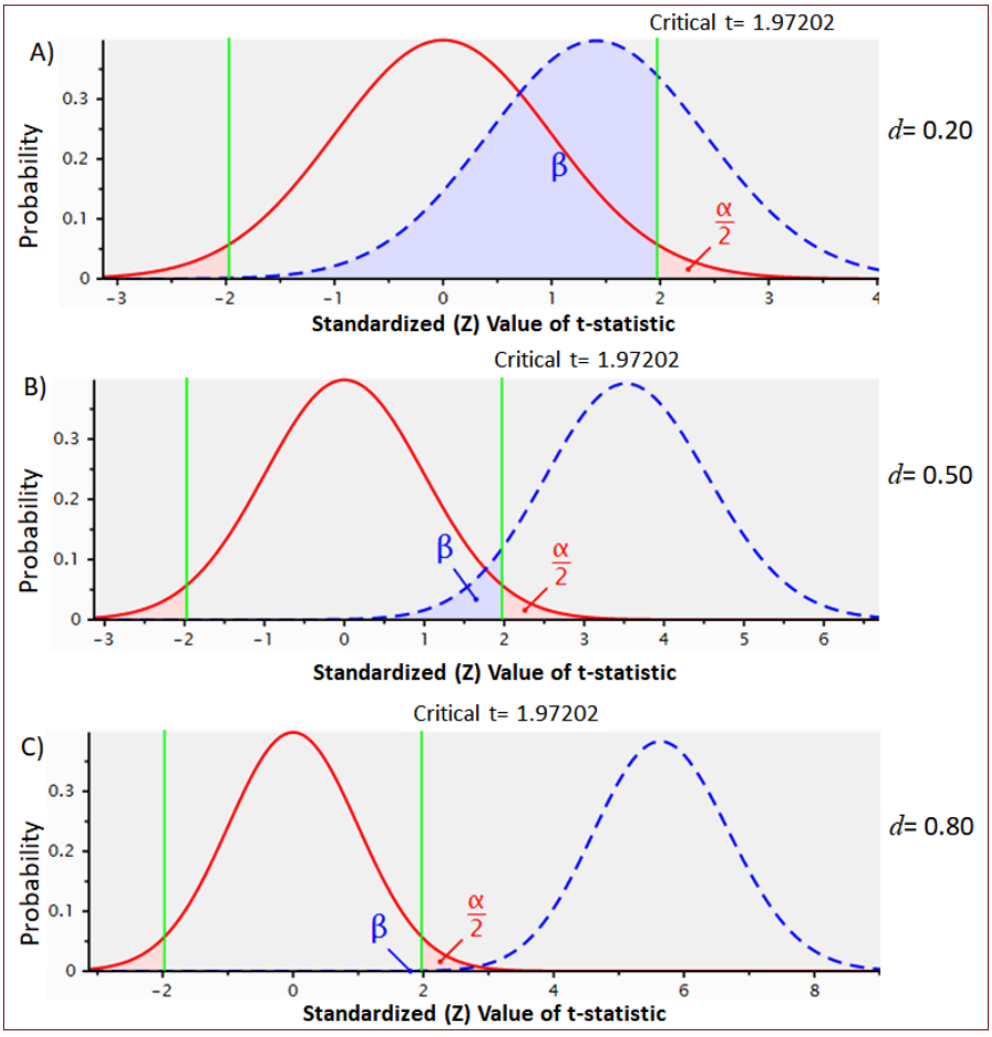

Figure 1. The relationship between Type I (α) and Type II (β) error in the following situation: n1=100, n2=100, α=0.05, t=1.97202, df=148, power (1-β)=0.80 for all. **A**) Effect size=0.20; **B**) Effect size=0.50; **C**) Effect size=0.80.

ES estimates should also be reported in conjunction with a confidence interval (*CI*) because an *ES* sample is a random variable. A large *CI* should be interpreted with caution due to imprecision [30]. A large *ES* and no statistical significance implies that sample size needs to be increased, while the opposite, a statistical significance (*P* value) in conjunction with a small *ES* implies that the result indicates that the significance only occurred due to the sample size increase (**Table 3**) [30,31]. Statistical errors can be better detected by reporting the *CI* and *ES* estimates (**Figure 2**) [20].

Table 3. Interpretation of results based on data variation (fictitious data)

Sample size (n)	Statistical significance	Effect size	Interpretation
small	not significant	large	no reliable conclusion (low statistical power, type II error)
appropriate	significant	large	reliable conclusion
appropriate	not significant	small	reliable conclusion
large (big data)	significant	small	no reliable conclusion (very high statistical power, type I error)

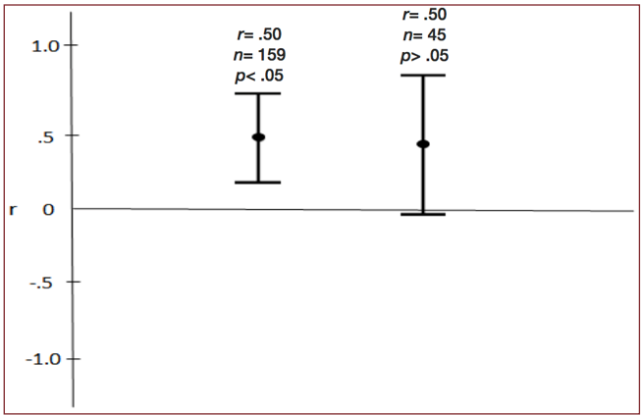

Figure 2. Confidence interval to support effect size interpretation (fictitious data).

CONCLUSION

ES is a measure involving the concept of clinical significance, while the *P* value involves that of statistical significance. Despite the fact that there are several methods for calculating *ES*, its major objectives are: 1) to validate the statistical significance test, and 2) to allow for a comparison of results from different studies with each other. Therefore, the combined reporting of *ES*, *CI*, and *P* value aims to enhance interpretation and prevent misinterpretation of data, and promotes clinical decision based on evidence-based studies.

REFERENCES

1. Cohen J. Statistical power analysis for the behavioral sciences. 2th ed. Ney Jersey, NJ: Lawrence Eribaum; 1988.
2. Sullivan GM, Feinn R. Using effect size – or why the *p* value is not enough. J Grad Med Educ 2012;4:279-282. https://doi.org/10.4300/JGME-D-12-00156.1
3. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485. https://doi.org/10.1136/bmj.311.7003.485
4. Cohen J. The earth is round (p<.05). Ame Psychol 1994;49:997-1003. https://doi.org/10.1037/0003-066X.49.12.997
5. Johnson D. The insignificance of statistical significance testing. J Wildl Manage 1999;63:763-72. https://doi.org/10.2307/3802789
6. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016;70:129-33. https://doi.org/10.1080/00031305.2016.1154108
7. Ferguson CJ. An (ES) primer: A guide for clinicians and researchers. Prof Psychol Res 2009;40:532-8. https://doi.org/10.1037/a0015808
8. Kirk RE. Practical significance: A concept whose time has come. Educ Psychol Meas1996;56:746-59. https://doi.org/10.1177/0013164496056005002
9. Pandis N. The effect size. Am J Orthod Dentofacial Orthop 2012;142:739-40. https://doi.org/10.1016/j.ajodo.2012.06.011
10. Espírito-Santo H, Daniel F. Calculating and reporting effect sizes on scientific papers (1): $p < 0.05$ limitations in the analysis of mean differences of two groups. RPICS 2015;1:3-16.
11. Khalilzadeh J, Tasci ADA. Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. Tour Manag 2017;62:89-96. https://doi.org/10.1016/j.tourman.2017.03.026
12. Ferrari R, Caram LMO, Garcia T, Paiva SAR, Vale AS, Tanni SE. Effect size. Pneumol Paul 2016;29:73-4.
13. Fritz CO, Morris PE, Richler JJ. (ES) estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012;141:2-18. https://doi.org/10.1037/a0024338
14. Nakagawa S, Cuthill IC. effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007;82:591-605. https://doi.org/10.1111/j.1469-185X.2007.00027.x
15. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Delineando a pesquisa clínica. 4th ed. Porto Alegre: Artmed; 2015.
16. Bakker M, Dijk AV, Wicherts JM. The rules of the game called psychological science. Perspect Psychol Sci 2012;7:543-54. https://doi.org/10.1177/1745691612459060
17. Anderson SF, Kelley K, Maxwell SE. Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. Psychol Sci 2017;28:1547-62. https://doi.org/10.1177/0956797617723724
18. Bradley MT, Brand A. Alpha values as a function of sample size, effect size, and power: accuracy over inference. Psychol Rep 2013;112:835-44. https://doi.org/10.2466/03.49.PR0.112.3.835-844
19. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. TSS 2014;1:19-25.
20. 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. https://doi.org/10.1007/s10654-016-0149-3
21. Hedges LV. Distributional theory for Glass´s estimator of effect size and related estimators. J Educ Behav Stat 1981;6:107-28. https://doi.org/10.3102/10769986006002107
22. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res 1976;5:3-8. https://doi.org/10.3102/0013189X005010003
23. Lalongo C. Understanding the effect size and its measures. Biochem Med 2016;26:150-63. https://doi.org/10.11613/BM.2016.015
24. Levine TR, Hullett CR. Eta squared, partial eta squared, and misreporting of effect size in communication research. Hum Commun Res 2002;28:612-25. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x
25. Bakeman R. Recommended effect size statistics for repeated measures designs. Behav Res Methods 2005;37:379-84. https://doi.org/10.3758/BF03192707
26. Espírito-Santo H, Daniel F. Calculating and reporting effect sizes on scientific papers (2): Guide to report the strength of relationships. RPICS 2017;1:53-64.
27. Maher JM, Markey, CM, Ebert-May D. The other half of the story: effect size analysis in quantitative research. CBE-Life Sci Educ 2013;12:345-51. https://doi.org/10.1187/cbe.13-04-0082
28. Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculation Cohen's f2, a measure of local effect size, from PROC MIXED. Front Psychol 2012;3:1-6. https://doi.org/10.3389/fpsyg.2012.00111
29. McHugh ML. The odds ratio: calculation, usage, and interpretation. Biochem Med 2009;19:120-6. https://doi.org/10.11613/BM.2009.011
30. Téllez A, García CH, Corral-Verdugo V. Effect size, confidence intervals and statistical power in psychological research. Psychol Russia 2015;8:27-47. https://doi.org/10.11621/pir.2015.0303
31. Nyirongo VB, Mukaka MM, Kalilani-Phiri LV. Statistical pitfalls in medical research. Malawi Med J 2008;20:15-18. https://doi.org/10.4314/mmj.v20i1.10949