Clinimetrics Corner: Use of Effect Sizes in Describing Data

Chad Cook PT, PhD, MBA, OCS, FAAOMPT

Recently, while working on the power analysis for a knee replacement, randomized controlled trial, we identified a study performed by Huang and colleagues[1] that used a similar design as our targeted trial. Like Huang et al's[1] study, we were interested in comparing a high flex knee total joint arthroplasty to a 'traditional' version. In their study, they tracked patient outcomes for 2 years using the Knee Society Score as an outcome measure. Using the previously published data, I calculated the effect size of the intervention and reported this figure to my colleagues. Their response: "what is an effect size?"

Effect size is a name given to a family of indices (currently there are over 40 types[2]) that measure the strength of a treatment effect[3,4]. In other words, effect size quantifies the true magnitude of the measured intervention, by providing a dedicated value (a numeric score) when comparing two (or more) groups[5]. For those who struggle with understanding the concept of *true magnitude* of an intervention, consider the following metaphoric example. A firecracker is an explosive device with a small effect size or magnitude. Dynamite is an explosive device with a larger effect size or magnitude. If creating a tunnel for a highway system is the intervention of choice for engineers, dynamite exhibits the strongest magnitude and demonstrates the strongest effect size. In rehabilitation, a theoretical comparison of two interventions is also possible. For example, for treatment of a rotator cuff surgical repair, a dedicated exercise program targeted to safely and steadily strengthen the rotator cuff mechanism is considered to have a larger effect size than the use of a moist hot pack.

ABSTRACT: Effect size is the name provided to a family of indices that are designed to measure magnitude of a targeted intervention. Generally, effect size measures are divided into three groups: 1) standardized difference measures 2) correlational measures, and 3) odds ratios. Effect sizes are valuable in indicating actual magnitude of selected interventions as the values are independent of sample size and statistical significance measures. Effect sizes are used to generate power requirements for a study and are the value of choice when calculating meta-analyses. Typically, effect size measures associate well with global rating of change scores for nearly all interventions.

KEYWORDS: Correlational Measures, Effect Size, Odds Ratio, Standardized Difference

Types of Effect Sizes

Effect sizes are generally divided into three groups: 1) standardized difference measures 2) correlational measures, and 3) odds ratios[6]. Standardized differences of two groups include Cohen's *d* and Hedges' *g*. Standardized measures of more than 2 groups (e.g. when an ANOVA is used) requires the use of Eta squared (h^2), partial Eta squared (h_p^2), omega squared (w^2), or the Intraclass correlation (r_I). Correlational effect sizes are used to measure the standardized covariance among two targeted variables. Odds ratios (with confidence intervals) are a form of effect size used during measurement of an inherently dichotomous construct[7]. Odds ratios measure the chance of success (or failure) in the experimental group as compared to the chance of success (or failure) in the control group.

How are Effect Sizes Calculated?

Effect sizes are calculated during comparison of experimental versus a control group of a different intervention (most common), experimental versus baseline interventions (least common and most potentially biased), or experimental versus a control group of no intervention. Although effect size can be calculated for all forms of measures, such as continuous, ordinal, or dichotomous, the mea-

Assistant Professor, Department of Surgery, Duke University Medical Center, Durham, NC.
Address all correspondence and requests for reprints to: Dr. Chad Cook, chad.cook@duke.edu

sure of standardized mean difference (Cohen's *d*) between two comparative groups that measure continuous data is the most common[3,8].

Calculating the standardized mean difference involves subtracting the mean of the intervention by the mean of the control group (which can involve no treatment, a control treatment, or the baseline measure) and dividing the within-group or pooled standard deviation of the two groups. A within-group or pooled standard deviation involves calculating the standard deviation for both groups at one time, versus two different standard deviations. The following equation outlines the calculation for Cohen's *d*.

$$\text{Cohen's } d = \frac{(\text{Mean intervention group}) - (\text{Mean control group})}{(\text{Standard Deviation})}$$

It's important to note that the outcome variable used in the effect size measure will dictate whether the effect size will be positive or negative. For example, for many outcome variables, such as the Roland Morris disability scale and a numeric pain scale measure, a lower value reflects a better outcome. Consequently, using the formula in a study where the intervention group improves more so than the control group will result in a negative calculated effect size. An understanding of the outcomes values is indeed necessary to recognize whether the effect size suggests the intervention is worse or better than the control group.

Because the calculated values provide little guidance in indicating the magnitude of an intervention, Cohen[9] created straitjacketed guidelines for effect size outcomes. His provision of intervals corresponding to trivial, small, medium, and large effect sizes, include: <0.20 (trivial), ≥0.20 to <0.50 (small), ≥0.50 to <0.80 (medium), and ≥0.80 (large). These values also correspond to percentages of the control group that fall below the mean values of those in the experimental group. For example, an effect size of 0.0 (trivial) suggests that approximately 50% of the control group would have outcomes that fall below the experimental group. An effect size of 0.4 (small) suggests that approximately 66% of the control group would have outcomes that fall below the experimental group. A large effect size of 0.80 indicates that 79% of the control group scored lower. Table 1 provides the percentage guidelines and equivalencies for effect size measures.

How are effect sizes used?

Effect size measures allow greater precision in determining the true magnitude of the intervention when results are not large or obvious (or statistically significant). Unlike significance tests, effect sizes are independent of sample size[3,4] and are useful singular measures when evaluating under- and overpowered studies. Statistical significance testing, which is often used as an outcome measure of the effectiveness of an intervention, provides substantial limitations in comparison to effect size. Studies with very large sample sizes and a small effect may still easily reach statistical significance, whereas conversely, studies with small sample sizes but strong effect size may fail to reach statistical significance[10].

Take the following example from the literature. In 2007 we reported that osteoarthritis (OA) affects self-report of mental health leading to a greater number of days reported where mental health was different in patients with OA versus without[11]. Although those with OA indicated 7.12 days where their mental health was not good (SD=10.56) and the control group of no-OA reported similar findings (7.20 days, SD=10.01) the finding was statistically significant ($p<0.01$). Statistical significance was met only because of the substantial sample size (6,172 subjects with OA and 31,523 subjects without OA); as the actual effect size was only 0.007. In contrast, consider a platform presentation recently accepted at the American Association of Hip and Knee Surgeons annual meeting by Krenzel et al[12]. In the study, one group (N=21) received posterior capsular injections of Ropivacaine after total knee replacement whereas a control group (N=20) received normal treatment with no injection. Between 8 to 48 hours post-operation, the group that received Ropivacaine improved in range of motion and pain although the findings were not statistically significant ($p=0.11$). Nonetheless, the effect size of

TABLE 1. Effect size strength and corresponding percentage guidelines.

Cohen's Interpretation of Effect Size Strength	Effect Size	Percentage of control group below average person in experimental group
Trivial	0.0	50%
	0.1	54%
Small	0.2	58%
	0.3	62%
	0.4	66%
Moderate	0.5	69%
	0.6	73%
	0.7	76%
Large	0.8	79%
	0.9	82%
	1.0	84%
	1.2	88%
	1.4	92%
	1.6	95%
	1.8	96%
	≥2.0	98% or greater

Ropivacaine injection was 1.12, suggesting that the clinical magnitude of the intervention was strong and that with an appropriate sample size, the study would have been statistically significant. Many manual therapy studies suffer from smaller sample sizes, which may impact the ability to truly identify differences between findings.

Effect size for a specific intervention has been said to be predetermined or, theoretically consistent among multiple studies[13] and is the currency used in meta-analyses[5]. Meta-analyses summarize the findings from a specific area of research and allow pooling of findings for measurement of a common effect size. Differences in effect size are typically reflective of study bias or design variation, more so than disparity in intervention outcome[14]. Variations include control groups, type of care provided, frequency of care, outcomes measures, and providers.

The most frequent use of effect size is during a pre-study and post-study power analyses. Although the power of a study is influenced by three components: 1) sample size, 2) the selected cutoff for significance, and 3) the effect size of the intervention, effect size remains the factor that may most economically and pragmatically improves the outcome of a comparative trial. Studies with interventions that exhibit small effect size will require substantially larger sample sizes and a less conservative cutoff for significance versus studies demonstrating larger or stronger effect size to demonstrate statistical significance[15].

Limitations of Effect Size Measures

One could argue than including effect size measures during reporting of findings provides substantially more information than commonly practiced actions such as publishing raw findings or stating whether the study was statistically significant. Nonetheless, there are instances where effect size values may provide misleading findings, specifically when bias is introduced within the study design, when data are not normally distributed or independent[16], and when standard deviations are so high that future comparisons of the study are not likely[17]. In addition, reported effect size values decline over time for a number of reasons[14]. Firstly, initial studies typically compare an intervention with no treatment or a true control. Effect size for the intervention is generally larger as the intervention is compared against no care. Follow-up studies are typically compared against the *accepted treatment of the time* and comparative means are more similar, thus the effect size actually is smaller. Furthermore, although effect size is independent of sample size, it is not independent of study design quality. As studies involving a dedicated idea mature, study designs improve and standard deviations decline. Study improvement, which involves careful design control, will also lead to a lower reported effect size.

Summary

It has been stated, that the greatest benefit of an effect size is the ability to translate a patient's change in health status to a standardized value (effect size)[18]. Effect sizes are theorized to correlate well to anchor based assessment of global health such as the global rating of change (GRoC)[19]. This is supported by the findings of Middel et al[20] who demonstrated that self-reported GRoC findings corresponded very well with effect size measures for patients receiving care for heart failure. A patient's GRoC reflects their own rating of the overall change in treatment in comparison to their initial starting point of care. Middel et al[20] found that a trivial effect equated well with the GRoC self report of *no change*, small effect correlated with self report of *a little better*, medium effect correlated with self report of *moderately better*, and large effect was somewhat concordant with self report of *a great deal better*. Nonetheless, it is important to note that clinically significant changes for some individuals and disease-related consequences may be .2 whereas for others may be .8[18].

The following websites provide Microsoft Excel-based effect size calculators (for Cohen's *d*) for ease of measurement.

Cohen's d http://web.uccs.edu/lbecker/Psy590/escalc3.htm

Cohen's d http://www.cemcentre.org/renderpage.asp?linkID=30325017

Cohen's d for t and f tests. http://davidmlane.com/hyperstat/effect_size.html

REFERENCES

1. Huang HT, Yuan S, Wang GJ. The early results of high-flex total knee arthroplasty. A minimum of 2 years of follow up. *J Arthoplast* 2005;20:674–679.
2. Kirk RE. Practical significance: A concept whose time has come. *Educational Psych Measurement* 1996;56:746–759.
3. Rosenthal R, Rosnow RL. *Essentials of Behavioral Research: Methods and Data Analysis*. 2nd ed. New York: McGraw Hill, 1991.
4. Rosnow RL, Rosenthal R. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Pyschological Methods* 1996;1:331–340.
5. Parker RI, Hagan-Burke S. Useful effect size interpretations for single case research. *Behavior Ther* 2007;38:95–105.
6. Rosnow RL, Rosenthal R. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Pyschological Methods* 1996;1:331–340.
7. McCartney K, Dearing E. Evaluating effect sizes in the policy arena. *Evaluation Exchange* 2002;8:1.
8. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969.
9. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.
10. Valentine JC, Cooper H. Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington DC; *What Works Clearinghouse*. 2003.
11. Cook C, Hegedus E, Pietrobon R. Osteoarthritis and the Impact on Quality of Life Health Indicators. *Rheumatol Internat* 2007;27:315–321.
12. Krenzel B, Cook C, Viens N, et al. Isolated Posterior Capsular Injections of Ropivacaine During Total Knee Arthroplasty: A prospective, randomized, double-blind,

placebo-controlled study of post-operative pain and function. *American Association of Hip and Knee Surgeons Annual Meeting*. November 7–9, 2008; Dallas, Tx.

13. McGraw KO, Wong SP. A common language effect-size statistic. *Psychological Bulletin* 1992;111:361–365.
14. Gehr BT, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Med Research Method* 2006;6:25.
15. Whitley E, Ball J. Statistics review 4: Sample sized calculations. *Crit Care* 2002;6:335–341.
16. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Sci* 2007;82:591–605.
17. Killeen P. An alternative to null-hypothesis significance testing. *Psychol Sci* 2005; 16: 345–353.
18. Kane R. *Understanding Health Outcomes Research*. Sudbury, MA: Jones & Bartlett, 2006.
19. Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. *Cont Clin Trials* 1989;10:407–415.
20. Middel B, Stewart R, Bouma J, van Sonderen E, van den Heuvel WJA. How to validate clinically important change in health-related functional status. Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating? *J Evaluation Clin Pract* 2001; 7:399–410.