Journal of Athletic Training 2016;51(12):1045–1048
doi: 10.4085/1062-6050-51.12.14

www.natajournals.org

short report

Statistical Primer for Athletic Trainers: Using Confidence Intervals and Effect Sizes to Evaluate Clinical Meaningfulness

Monica Lininger, PhD, LAT, ATC*; Bryan L. Riemann, PhD, ATC, FNATA†

*Athletic Training Education Program, Northern Arizona University, Flagstaff; †Biodynamics and Human Performance Center, Armstrong State University, Savannah, GA

Objective: To describe confidence intervals (CIs) and effect sizes and provide practical examples to assist clinicians in assessing clinical meaningfulness.

Background: As discussed in our first article in 2015, which addressed the difference between statistical significance and clinical meaningfulness, evaluating the clinical meaningfulness of a research study remains a challenge to many readers. In this paper, we will build on this topic by examining CIs and effect sizes.

Description: A CI is a range estimated from sample data (the data we collect) that is likely to include the population parameter (value) of interest. Conceptually, this constitutes the lower and upper limits of the sample data, which would likely include, for example, the mean from the unknown population. An effect size is the magnitude of difference between 2 means. When a statistically significant difference exists between 2 means, effect size is used to describe how large or small that difference actually is. Confidence intervals and effect sizes enhance the practical interpretation of research results.

Recommendations: Along with statistical significance, the CI and effect size can assist practitioners in better understanding the clinical meaningfulness of a research study.

Key Words: statistics, reporting statistical findings, data interpretation

This is the second paper in our series seeking to facilitate clinicians' understanding of statistical results. In the first paper,[1] we discussed the difference between *statistical significance*, which reflects the influence of chance on a study's outcome, and *clinical meaningfulness*, which indicates whether the results have relevance to athletic training practice. We will now review confidence intervals (CIs) and effect sizes to assist clinicians in assessing clinical meaningfulness.

PREVIOUS EXAMPLE

Our previous paper presented a hypothetical study of 40 randomly selected healthy and physically active participants with restricted ankle dorsiflexion.[1] The participants were randomly assigned to 1 of 2 groups: control (standard stretching) or experimental (myofascial release followed by standard stretching). Ankle-dorsiflexion range of motion (ROM) was measured using a standard goniometer before and after the intervention program. After the 3-week program, ROM improvements (mean ± standard deviation) were 5.7° ± 1.7° and 6.8° ± 1.5° for the control and experimental groups, respectively. Based on the results of a statistical test comparing the groups ($t_{38} = 2.05$, $P = .047$) and the computed P value that was less than the α level of .05, the researchers claimed that myofascial release with stretching resulted in greater ROM improvement. Inherent to this interpretation is that the 1.1° difference in ROM exceeded what would be expected if ROM improvements did not differ between the groups.

DEFINITIONS

Most often, statistical analyses are conducted with a null hypothesis stating that there is no difference between groups or across time.[1] Once statistical significance is identified, instead of simply stating that the null hypothesis was rejected or not, we can calculate and interpret CIs and effect sizes. These values will help us determine the clinical meaningfulness of the 1.1° difference in ROM found between the groups. As described in our first paper, because research relies on samples selected from populations, our estimate about what exists in the entire population might differ slightly from the actual value in the population and may also fluctuate among repeated studies pulling samples from the same target population. By providing a range of values, CIs offer a margin of error around the computed statistics (ie, the difference in ROM improvements between the groups) that likely captures the true treatment difference that would have been identified if the entire population had been studied. Confidence intervals provide information regarding the location and precision of potential values (means, difference between 2 means, etc) had the entire population been studied. By definition, a *confidence interval* is a "range of values within which the population value is believed to lie at a given level of statistical confidence."[2(p108)] Statistical confidence is related to a selected level of statistical significance (α). Research relies on selecting a sample from the population and using the sample to infer what could be expected if the entire population had been studied. *Statistical confidence* reflects

the percentage of repeated samples that would indicate the true effect in the population. When used with widely understood measures, such as joint laxity, muscle strength, and ROM, CIs provide the clinician with more information that can be incorporated into clinical decision making (eg, risk benefit, cost benefit).

A standardized *effect size* is the magnitude of difference between 2 sample means expressed on a unitless scale.[3] In other words, the effect size is a measure of the difference between 2 means while accounting for the differences in responses within the sample. Unlike statistical significance and CIs, effect sizes have the advantage of not being heavily influenced by sample sizes. Additionally, as we will discuss, their magnitude is independent of the units associated with the measure. Thus, when the measures are not as widely understood—such as measures of balance, hydration (urine specific gravity), and fatigue—computation and interpretation of standardized effect sizes can be used to assess clinical meaningfulness.

UNDERSTANDING CONFIDENCE INTERVALS

Confidence intervals provide a range of possible values when using samples to estimate populations. For the practitioner, they answer 2 questions: "What is the magnitude of the effects (ie, differences)?" and "How precise does the study estimate the effect to be?" The CI (precision) directly influences its clinical utility. An intervention that results in a reported 95% CI of 1° to 15° for ankle ROM improvement does little to assist the practitioner in making clinical decisions. Differences of 5° or more might be deemed important, so perhaps the intervention should be recommended. Conversely, differences of 1° to 2° are likely trivial. In this case, the clinician is left to await more data from which to make decisions with a greater assurance of patient benefit. The true population effect might be a difference of as little as 1° or as much as 15°; this lack of precision leads to uncertainty about the actual ROM improvement. In contrast, a 95% CI of 5° to 8° provides a more precise estimate of the true population effect. This can be interpreted to indicate that we are 95% sure that the true population estimate is between 5° and 8° of ROM improvement. Consequentially, the clinician can have more certainty about the actual change in ROM due to the intervention when the CI is narrower. In that case, the clinician is better positioned to weigh the risks and costs of care with confidence that most patients will likely benefit from a treatment. For example, deciding on whether the time required to implement the ROM intervention is justified is easier when the 95% CI is 5° to 8° versus 1° to 15° (the latter reflects less confidence about the benefit of implementing the intervention).

However, a ROM improvement that falls within the CI is not guaranteed for every patient who matches the study sample characteristics and receives the intervention. Furthermore, although substantial statistical evidence (95%) suggests we can expect the true improvement in the population to fall within the CI, there is a possibility (ie, role of chance = 5%) that the true effect is outside the reported CI. Similar to our discussion on statistical significance and type I statistical errors,[1] we cannot determine if the CI computed from a single study has captured the true population effect or not. Most often, 95% CIs coincide with the common use of $\alpha = .05$ to define statistical significance, but there are occasions when various degrees of certitude might be appropriate. When a higher level of confidence is needed, 99% CIs, which correspond to $\alpha = .01$, may be used, as in studies by the US Food and Drug Administration that involve potentially severe side effects. When a situation calls for less confidence, 90% CIs, which correspond to $\alpha = .10$, may be used, as in pilot testing of an intervention before conducting a large-scale study.

As mentioned earlier, the width (precision) of CIs directly affects clinical utility. Factors that influence the width of the CI include sample size, variability of the data (varying responses among individuals in the study), and the level of confidence. Larger sample sizes more likely better represent the population, and the resulting CI widths are narrower. Variability within the sample indicates that the treatment effect was not consistent among study participants, which can occur when inclusion or exclusion criteria are broad (eg, participants with a wide variety of ankle ROMs are enrolled in the study). Finally, higher levels of confidence (99% versus 95%) produce wider CIs.

Using the data from our hypothetical example concerning whether myofascial release augments the effects of a 3-week ankle-dorsiflexion stretching program, the 95% CI for the difference in ROM improvements is 0.07° to 2.13°. Therefore, in our sample, we found a mean difference of 1.1°, and we are 95% confident that the actual difference in the whole population is between 0.07° and 2.13°. Given the additional 5 minutes needed to perform myofascial release, we would hardly consider the difference between the groups to be clinically meaningful. Furthermore, CIs must also be interpreted in terms of the reproducibility of the measurements. Often, the *minimal detectable change* statistic (a topic for a future paper in this series) is used to provide a minimal threshold value that 2 subsequent assessments must exceed to be assured the change is true and not the result of random measurement error. The minimal detectable change for active and passive goniometric dorsiflexion measurements has been reported to be between 5.7° and 7.4°,[4] so we would conclude that the myofascial release in conjunction with the dorsiflexion strength intervention does not produce clinical meaningfulness. Thus, although statistical significance is present, adding myofascial release before dorsiflexion ROM does not result in clinical meaningfulness.

Finally, it is worth noting that CIs can also be used to assess statistical significance. Unlike *P* values, which can only indicate statistical significance, CIs computed for the differences between groups in an intervention study can address both statistical significance and the magnitude of change for assessment of clinical meaningfulness. In our example, because the 95% CI for the difference between the groups did not contain zero (which would indicate no difference between groups), we can conclude that the groups are indeed statistically significantly different (ie, they exceed chance expectation) at an α level of .05 (actual *P* value = .047).

UNDERSTANDING EFFECT SIZES

When we find a statistically significant difference between 2 means, as in our example, standardized effect sizes can help us to further understand the size of the

Table. Effect-Size Interpretation Conventions

Effect Size	Cohen Convention[5]	Rhea Convention[6]	Percentage of Control Group Below the Result of the Average Experimental Group Participant
0.20	Small		58
0.35		Upper limit of trivial	64
0.50	Medium		69
0.80	Large	Upper limit of small	79
1.5		Upper limit of moderate	93

differences attained. As we will describe, reporting an effect size informs the consideration of clinical meaningfulness; however, the effect size should be presented with other statistical findings, such as the associated *P* value. The *P* value shows that there is a difference exceeding your pre-identified level of chance, but an effect size quantifies the size of the difference or strength of the relationship identified by the *P* value. The presentation of the effect size allows the practitioner to answer the question, "How large (or small) a difference did the intervention produce between the treatment and control groups?"

When considering the difference between 2 samples of scores, either between 2 groups or the same group assessed twice, *standardized effect sizes* are most often computed as the difference between the 2 sets of scores divided by a standard deviation. Depending upon the research design, the standard deviation may be from the control group or pretest scores, or when the research design does not include a control group, the pooled (combined) standard deviation of the 2 groups. Computed in this manner, standardized effect sizes become equivalent to standardized scores (z scores) and the standard normal distribution described in most introductory statistics and research textbooks. Standardized effect sizes computed using this method can be interpreted in 2 ways. The first interpretation is by how many standard deviations the average person in the experimental group differs from the average person in the control group. For example, an effect size of 0.4 would indicate that the experimental group is, on average, 0.4 standard deviations greater than the control group. An additional way of interpreting effect sizes is to use the standard normal distribution to determine the amount of overlap between the 2 distributions of scores. An effect size of 0.4 would indicate that the average person in the experimental group has a higher score than 66% of those in the control group.

Quantifying the standardized effect difference between the groups is useful, but caution is needed when applying the information clinically. Cohen[5] provided a template on which to judge the clinical meanings of effect sizes in psychology research and suggested that an effect size of 0.2 is *small*, 0.5 is *medium*, and 0.8 is *large*. Although this is helpful, these conventions were developed for psychology research, not athletic training research. In psychology research, small effect sizes are expected. Rhea[6] proposed that conventions for strength training research be modified to less than 0.35 as *trivial*, 0.35 to 0.80 as *small*, 0.80 to 1.50 as *moderate*, and greater than 1.5 as *large*. The Table summarizes these effect-size interpretation conventions. For athletic training research, which encompasses a range of disciplines that include psychological and strength training research, defining universal effect-size boundaries is not possible. Rather, equipped with an understanding of standardized effect sizes, the clinician can use the reported effect size to roughly estimate the clinical meaningfulness of the difference reported between the groups. Finally, when assessing the clinical meaningfulness of standardized effect sizes, the clinician must consider the characteristics of the populations studied, particularly the inclusion and exclusion criteria, and be certain that reliable outcome measures were used.

An additional advantage of standardized effect sizes is that studies using different outcome measures can be compared. For example, if 2 studies examining a muscle-strengthening intervention used different outcome measures (eg, isokinetic peak torque versus 1-repetition maximum), we could determine if the 2 studies identified similar increases in muscle strength by comparing the reported effect sizes. The calculation of an effect size takes into account the different scales used to measure muscle strength in the 2 studies (Newton-meters versus kilograms), allowing the results to be compared. Furthermore, the publication of effect sizes allows researchers to perform meta-analyses. Previously reported effect sizes help investigators performing an a priori power analysis to determine the needed sample size. The concept of a power analysis will be presented in a forthcoming paper in this short series.

From our hypothetical research study examining myofascial release, the effect size of –0.65 suggests that, on average, the ROM change for participants in the control group was 0.65 standard deviations less than for the participants in the experimental group. (See the Appendix for further detail on how this effect size was calculated.) In other words, on average, myofascial release before stretching improved ROM by 0.65 standard deviations more than stretching alone. According to the Cohen categories,[5] we could describe the intervention as having a medium effect of 0.5, which "is visible with the naked eye of a careful observer."[7(p156)] This description suggests that the intervention of myofascial release followed by stretching resulted in an ankle ROM improvement that a clinician could differentiate without conducting statistical analyses. However, using the Rhea criteria,[6] this effect would be considered small. In this example, the Rhea interpretation of effect size is more consistent with the conclusions drawn from examining the 95% CI around the mean difference.

CONCLUSIONS

Reporting CIs and effect sizes is important due to the additional practical information provided. A CI is a range estimated from sample data that has the predetermined likelihood of including the population parameter of interest. Conceptually, the lower and upper limits of the sample data would likely include, for instance, the mean from the unknown population. An effect size is the magnitude of difference between 2 means. When 2 means are statistically different, an effect size is used to interpret the size of the difference in relation to the sample distribution.

RECOMMENDATIONS

If we had presented only the statistical results of our hypothetical ankle-stretching study ($t_{38} = 2.05$, $P = .047$), then readers could conclude only that the 2 groups were significantly different but nothing further. By including both the 95% CI (0.07°, 2.13°) and the effect size (–0.65), we can now say that the groups were indeed statistically different (control = 5.7° ± 1.7°, experimental = 6.8° ± 1.5°), that the population mean difference was between 0.07° and 2.13°, and that the intervention had a medium effect when comparing the scores of participants in the control group with those of the treatment group. These additional statistics should be reported by investigators and interpreted by consumers in their critical appraisal of clinical research.

APPENDIX

To find the practical meaning of the differences between 2 groups, such as the ROM intervention example in this paper, an effect size can be calculated using the following information:

$$d = \frac{\bar{x}_{control} - \bar{x}_{exp}}{s},$$

where $\bar{x}_{control}$ is the mean of the control group, $\bar{x}_{exp}$ is the mean of the experiment group, and s is the standard deviation. It should be noted that the standard deviation in the denominator (the standardizer) will change depending on the study design. Usually, the standard deviation associated with the control group (or the pretest for a within-subject comparison) is used unless there is no control group, in which case the pooled (weighted average) standard deviation across both groups is used:

$$d = \frac{5.7 - 6.8}{1.7} = -0.65.$$

REFERENCES

1. Riemann BL, Lininger M. Statistical primer for athletic trainers: the difference between statistical and clinical meaningfulness. *J Athl Train*. 2015;50(12):1223–1225.
2. Armstrong LE, Kraemer WJ. *ACSM's Research Methods*. Philadelphia, PA: Wolters Kluwer; 2015.
3. Lomax R, Hahs-Vaughn D. *An Introduction to Statistical Concepts*. 3rd ed. New York, NY: Routledge Taylor & Francis Group; 2012.
4. Krause DA, Cloud BA, Forster LA, Schrank JA, Hollman JH. Measurement of ankle dorsiflexion: a comparison of active and passive techniques in multiple positions. *J Sport Rehabil*. 2011;20(3):333–344.
5. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:274–288.
6. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. *J Strength Cond Res*. 2004;18(4):918–920.
7. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.

Address correspondence to Monica Lininger, PhD, LAT, ATC, Athletic Training Education Program, Northern Arizona University, PO Box 15094, Flagstaff, AZ 86011. Address e-mail to monica.lininger@nau.edu.