Principles of Statistics

What the Sports Medicine Professional Needs to Know

Bryan L. Riemann, PhD, ATC, FNATA[a,*],
Monica R. Lininger, PhD, LAT, ATC[b]

KEYWORDS

• Clinical meaningfulness • *P* values • Data interpretation • Confidence intervals
• Effect sizes • Statistical power • Minimal detectable difference
• Minimal important difference

KEY POINTS

- Statistical significance reflects the influence of chance on the outcome, whereas clinical meaningfulness reflects the degree to which the study results reported are relevant to sports medicine practice.
- When statistically significant differences are revealed, confidence intervals and effect sizes can be used to enhance the practical interpretation of the research results.
- Absolute reliability characteristics, such as the minimal detectable change, determine the extent of error around a measurement and, when coupled with an appropriate minimal important difference estimate, can assist in triangulating clinically meaningful changes in patients undergoing treatment.
- In the circumstance of nonstatistically significant results, evaluation needs to occur to determine if the study had adequate statistical power prior to concluding no difference or association exists.

Along with many disciplines in medicine and allied health, the evidence-based practice (EBP) movement has prompted practitioners in the field of sports medicine to have a better competency in understanding research. As new procedures, methods, and understanding are studied with the results presented in research studies, practicing sports medicine professionals are faced with evaluating both statistically significant and clinically meaningful benefit along with whether the results are pertinent to their patients. Unquestionably, interpreting statistical findings as part of the research

Disclosure Statement: No conflicts of interest pertaining to this work.
[a] Sports Medicine, Biodynamics and Human Performance Center, Georgia Southern University, 11935 Abercorn Street, Savannah, GA 31419, USA; [b] Physical Therapy and Athletic Training, Athletic Training Education Program, Northern Arizona University, PO Box 15094, Flagstaff, AZ 86011, USA
* Corresponding author.
E-mail address: briemann@georgiasouthern.edu

Clin Sports Med 37 (2018) 375–386
https://doi.org/10.1016/j.csm.2018.03.004
sportsmed.theclinics.com

evaluation process can be a daunting challenge for many practitioners. Few practitioners enter a field, such as sports medicine, wanting to develop and possess an extensive expertise in statistics. Fortunately, once command over some of the nomenclature and few key concepts, along with understanding the more common statistical procedures, is attained, practitioners can often begin to evaluate research. The purpose of this article is to provide minimal essentials a sports medicine professional needs to know about interpreting statistics and research results to facilitate the incorporation of the latest evidence into practice. Topics covered include the difference between statistical significance and clinical meaningfulness; effect sizes and confidence intervals; reliability statistics, including the minimal detectable difference (MDD) and minimal important difference (MID); and statistical power. To begin the discussion, some of the common research and statistical terms are presented in **Table 1**.

ROLE OF STATISTICS IN THE RESEARCH PROCESS

Research, the process of acquiring new knowledge through systematic data collection procedures followed by controlled and critical analysis of the data, is one of the foundations of EBP. It is the analysis of data part of this definition for which statistics become necessary. One of the first tasks after data are collected is to summarize the data so they may be reduced into smaller and more interpretable chunks. Descriptive statistics are those indicators that are used to portray the data in more interpretable chunks. The challenge is that no single descriptive statistic can represent an entire data set as a single value. For example, the mean change in range of motion can be described for 2 groups as 10°; however, closer inspection of the individual participants may yield a different perspective (**Table 2**). Thus, at minimum, the optimal approach is to provide some indicator of the data set center and the extent of variation around the center. The measurement scale of the data dictates the appropriate descriptive statistics to use (**Table 3**).

In research, it is not feasible to study all members of a defined population. Because research involves selecting a sample from a target population, assumptions and hypotheses must be examined to determine whether the obtained results are tenable or if they could have simply occurred due to chance. Additionally, a researcher often wants to generalize or make predictions about the population from which a sample was obtained. A population is defined as all individuals who meet the inclusion and exclusion criteria for a specific study. When a sample is drawn from the population, sampling error is likely incurred because it is probable that the sample does not perfectly represent (ie, duplicate) a population. Thus, estimates about what exists in the entire population based on the sample differ from the true reality. The magnitude of sampling error is attempted to be decreased through using optimal research design elements, including random sampling and random allocation (assignment), sufficiently sized samples, and reliable outcome measures. Despite the efforts to decrease sampling error, some uncertainty always exists; the challenge for researchers and practitioners alike is how to evaluate whether the difference/relationship results are real versus the likelihood that they occurred based on chance (ie, sampling error). By providing a *P* value, inferential statistics attempt to provide some indication regarding the extent to which chance might explain the results.

INTERPRETING *P* VALUES

Interpretation of the resulting *P* values from inferential statistics is an area of frequent confusion. Two perspectives on using *P* values have been described, significance testing and hypothesis testing. Over the years the 2 perspectives have often become

Table 1
Commonly used terms in research and statistics

Term	Definition
Population	The target population is all participants who meet a specified set of inclusion and exclusion criteria. In practice, it is not possible to have access to all persons meeting the criteria; the accessible population is the portion of the target population that the researcher has access from which to select participants.
Sample	The selected sample is the group of persons selected from the accessible population who are asked to participate in the study. The actual sample is the group of participants who complete the study and whose data is used for analysis.
Random sampling	The selection of selecting participants so that each has an equal chance of being chosen (reduce sampling bias). This method is needed to ensure the accuracy of inferential statistical interpretation.
Random allocation	Method of assigning participants to groups using randomization (ie, flip a coin or random number generator) so each participant has an equal opportunity to be assigned to a given group. By using, greater confidence exists that the groups are equivalent at baseline. Also referred to as random assignment.
Independent variable	The variable the researcher is manipulating (active) or selecting (attribute) to determine the effect on the dependent variable
Dependent variable	The outcome variable of interest in a study; in research, the approach is taken that the independent variable influences the dependent variable.
Descriptive statistics	Statistics that summarize the characteristics of the data that include describing the center of the data (ie, mean and median) and variation (ie, standard deviation and range)
Inferential statistics	Statistics that are used to generalize (infer) the results of a study sample to the wider population
Type I error	The error committed when the data demonstrate a statistically significant result although no true difference or association exists in the population
Type II error	The error committed when the data do not demonstrate a statistically significance result when a true difference or association exists in the population
Alpha (α)	The probability, ideally established before inferential statistical analysis is conducted, that a type I statistical error will be permitted. Most frequently .05 is used.
Beta (β)	The probability, ideally established before inferential statistical analysis is conducted, that a type II statistical error will be permitted. Most frequently .20 is used.
Confidence interval	A range of values calculated from the sample data, which are likely to contain the value for the population to a given level of confidence. Most often 95% CI is used. The width of a confidence interval provides an indication of the precision of estimated population value, thereby assisting with determining clinical meaningfulness of statistical test results.

(*continued on next page*)

Table 1
(*continued*)

Term	Definition
Effect size	The observed or expected change or association between the 2 or more variables. Reporting standardized effect sizes (ie, *d* family or *r* family) can assist with determining clinical meaningfulness of statistical test results.
Measurement validity	The accuracy of a measure to quantify what it is designed to measure
Measurement error	The difference between the true value of a quality and what is measured
Measurement reliability	The quality of a measure to be consistent and free from measurement error
Minimal detectable difference	The minimal quantity of change that exceeds measurement error
Minimal important difference	Quantity of change that a patient perceives as worthwhile
Statistical power	The probability a statistical test will correctly identify a difference or association that truly exists (ie, not commit a type II statistical error)

combined, which has led to much of the *P* value interpretation confusion as well as many challenges to using *P* values to interpret the results of a research study.[1–3] Practitioners do not need a full understanding of the *P* value debate among statisticians. Rather, practitioners need to appreciate what a *P* value from an inferential statistic indicates and how to combine that information with other tools, such as confidence intervals and effect sizes, to evaluate the clinical meaningfulness of a research result.

In current practice, most research using inferential statistics with *P* values is done as a null hypothesis significance test (NHST). In conducting the inferential statistical test, a null hypothesis is created. A null hypothesis is frequently stated as there is no difference between groups (ie, the invention is not effective) or no relationship exists. It is critical to understand that the *P* value is a conditional probability; assuming a true null hypothesis, the *P* value provides the probability that the differences or relationships yielded by the sample data are attributable to sampling error. Although the *P* value was originally described as the magnitude of evidence against the null hypothesis (significance testing), in current practice the *P* value is often compared

Table 2
The average range of motion improvement for the 2 groups is identical; however, inspection of the individual responses demonstrates variability that is characterized by the standard deviation

	Range of Motion Improvement (°)	
Participant Number	**Group 1**	**Group 2**
1	8	5
2	10	18
3	13	2
4	10	15
5	9	10
Average ± SD	10 ± 1.9	10 ± 6.7

Table 3
Measurement scales and descriptive statistics

Measurement Scale	Definition	Examples	Measure of Center	Measure of Spread
Nominal	Unordered categories	Ethnicity, gender, injury status	Mode	Number of categories containing values
Ordinal	Ordered categories	Manual muscle test grades, reflex grades	Median	Interquartile range
Interval	Equal intervals between scores but no true zero	Temperature, time on a 12-h clock, GRE score	Mean[a]	Standard deviation[a]
Ratio	Equal intervals between scores and true zero	Body weight, blood pressure	Mean[a]	Standard deviation[a]

Abbreviation: GRE, graduate record examination.
[a] In cases of data not normally distributed or influential outliers, the median and interquartile range are more appropriate.

against a threshold point (α level) to make a binary decision to either reject or fail to reject the null hypothesis (NHST). When the computed *P* value is smaller than α, statistical significance, or rejection of the null hypothesis, is concluded. The α level is established by the researcher early in the research process; most often .05 is used, although there are circumstances where different α levels may be more appropriate.[4]

The process of making a yes-or-no decision regarding the null hypothesis by comparing the *P* value against α raises the possibility for making 2 kinds of errors. The first, referred to as a type I statistical error, is when the computed *P* value causes the researcher to reject a null hypothesis that was really true (ie, false positive). In this circumstance, the researcher is concluding that there is a difference or relationship when in reality there is not. The probability of making a type I error is related to the α level. The second type of statistical error, a type II error, occurs when the *P* value prompts a researcher to erroneously conclude there is no difference or relationship, thereby failing to reject a false null hypothesis (ie, false negative). The probability of making a type II statistical error is referred to as β. It is important to recognize that α and β do not indicate the likelihood that a single study has led to error but rather the likelihood over multiple replications of the research.[1]

To illustrate NHST and *P* value interpretation, consider a hypothetical study in which 2 different terminal-phase anterior cruciate ligament rehabilitation programs, conventional and accelerated, are compared for improving single-leg hop test limb symmetry index (LSI). In conducting the study, a random sample of 40 individuals who underwent anterior cruciate ligament reconstruction is randomly assigned to 1 of the 2 rehabilitation programs. After the 4-week terminal phase of rehabilitation, the participants' single-leg hop test LSI (uninjured/injured $\times$ 100) is determined. After the program, the LSI values are 87.3% $\pm$ 6.2% and 91.9% $\pm$ 6.5% for the conventional and accelerated groups, respectively. Based on the results of a statistical comparison between the groups ($t_{38} = 2.28$, $P = .028$), the researchers claim that the accelerated program is better than the conventional program because the computed *P* value was less than .05. The core of this interpretation is that the 4.6% LSI difference between the 2 rehabilitation groups exceeded what would be expected if no difference existed in the LSI improvements between the 2 groups.

Despite the common practice of NHST, as described previously, there are a few additional aspects that warrant consideration. First, despite the attractiveness of making a dichotomous decision (yes or no) regarding the null hypothesis, there is little logic in considering the findings of a study reporting a *P* value of .048 as evidence supporting an intervention whereas a *P* value of .052 leads to the conclusion that an intervention did not work because it was slightly larger than alpha, the statistically significant threshold. The *P* value is a function of several factors, including the sample size and the effect size. An effect size is the magnitude of the difference/change in relationship compared with the variability between the study participants (discussed later). Thus, in addition to considering the effect size, prudent practice also includes considering the *P* value in conjunction with the sample size.[5] Although a difference between groups may not reach statistical significance with a sample size of 10, the same difference becomes statistically significant with a sample size of 20 (**Table 4**).

STATISTICAL SIGNIFICANCE DOES NOT MEAN CLINICAL MEANINGFULNESS

Whereas statistical significance provides an indication of the role of chance on the outcome, clinical meaningfulness is the importance and relevance of the result to sports medicine practice. Demonstrating a significant improvement in an outcome measure (eg, pain) because of a clinical intervention (eg, rehabilitation) is worthwhile; however, if the change is very small and took 6 weeks of 3 rehabilitation sessions per week, it may be concluded that it does not have clinical meaningfulness. Thus, clinical meaningfulness is a subjective decision that weighs the risks and costs (eg, time and money) with the benefits yielded by a study that attained statistical significance. Additionally, when determining clinical meaningfulness, the reliability of the measurement tools and the types of participants used in the study also need to be considered. Although unreliable outcome measures decrease the likelihood of reaching statistical significance, if the changes deemed statistically significant do not exceed measurement error or what a patient perceives as beneficial, they may have little clinical meaningfulness. The MDD and MID are 2 tools to assist with this decision (discussed later). The characteristics of the study participants also influence the magnitude of change; smaller changes likely occur with interventions that use healthier participants versus more severely injured patients. Thus, in assessing clinical meaningfulness, a practitioner needs to evaluate the stated inclusion/exclusion criteria and the final characteristics of the participants, the relevance and reliability of the outcome measures, and the magnitude of the changes/differences revealed by the investigation.

Table 4
Effects of sample sizes on statistical significance and effect sizes

Sample Size	Mean Limb Symmetry Index Difference Between Groups (%)	*t* Statistic	*P* Value	Statistically Significant at .05?	Cohen's *d* Effect Size
10 per group	4.6	1.61	.116	No	.72
20 per group	4.6	2.28	.028	Yes	.72
30 per group	4.6	2.79	.008	Yes	.72

Unlike the *P* value, sample size does not influence the unstandardized (mean difference) or standardized (Cohen's *d*) effect sizes.

EFFECT SIZES AND CONFIDENCE INTERVALS AS MEASURES OF CLINICAL MEANINGFULNESS

Again, after a statistically significant result is determined, it is important to decide how clinically meaningful it is. One way to do this is using an effect size that estimates the magnitude of the changes or associations revealed by the statistical results. There are 2 main types of effect sizes: unstandardized and standardized. The unstandardized effect size is simply the mean difference of the outcome measures used within 1 study. Using the single-leg hop test, the unstandardized effect size is 4.6%, because this is the mean difference between the 2 groups (see **Table 4**). Although an unstandardized effect size can be interpreted directly, it cannot be compared across multiple studies. The standardized effect size is expressed on a unitless scale and, therefor,e can be used in a larger scope. The *d* family (difference between groups) and *r* family (association or relationship between variables) are 2 types of standardized effect sizes.

Although there are more than 70 different effect size indexes,[6] some of the most common in each of the families are focused on for this article. Cohen's *d* is the most common effect size in the *d* family. This effect size index is the mean difference between the 2 groups divided by the standard deviation (a standardizer). The standardizer is what places this type of effect size on a unitless scale. Researchers commonly use 3 categories of small (.2), medium (.5), and large (.8) when describing the effect size. A Cohen's *d* effect size of .72 (see **Table 4**) is computed using the single-leg hop test LSI data of 91.9% $\pm$ 6.5% (accelerated) and 87.3% $\pm$ 6.2% (conventional). The mean difference of 4.6 is divided by the pooled SD[7] of 6.38. Therefore, the individuals in the accelerated rehabilitation protocol had a higher LSI by .72 SDs than those in the conventional protocol. According to the categories presented by Cohen,[8] this represents a medium effect between the 2 rehabilitation programs. Cohen first suggested these classifications for psychology research where smaller effect sizes are more common than in sports medicine research. More recently, Rhea[9] suggested classifications for strength training research of less than .35 as trivial, .35 to .80 as small, .80 to 1.50 as moderate, and greater than 1.5 as large. Another common effect size in the *d* family is an odds ratio. If a clinician is interested in the association of the presence of reinjury with those individuals who completed the accelerated protocol compared with those in the conventional protocol, an odds ratio could be calculated. For example, with an odds ratio of .91, the individuals in each rehabilitation group have approximately equal odds of reinjury. An odds ratio of 1.00 is exactly equal odds.

When describing the practical importance using indexes from the *r* family, the more commonly seen options include a form of η^2 or a Pearson product moment correlation coefficient. η^2 is the proportion of the variation in the dependent variable explained by the independent variable. η^2 can be positively biased, especially for samples with fewer than 30 participants. An alternative is the partial eta squared (η^2_p), which accounts for the variation in the dependent variable and the associated independent variable of interest in isolation. The Pearson product moment correlation coefficient describes the relationship between 2 continuous variables. As an effect size, it defines the strength of the relationship. Bilateral strength imbalance (newton meter), specifically in the quadriceps, may be related with the LSI. Using a Pearson product moment correlation coefficient quantifies this relationship. For example, a finding of .89 suggests that there is a strong, positive correlation between these 2 measures.

Another way to determine clinical meaningfulness is with confidence intervals, which provide a range of possible values drawn from samples to estimate the

population. With the example of the LSI scores after the 2 rehabilitation programs, the computed 95% CI suggests that the true population effect is between .51% and 8.67%. Confidence intervals can be considered at any level (eg, 90% or 99%), but the most commonly seen is 95%. This relates to the commonly seen use of $\alpha = .05$ as a determination of statistical significance. Similar to NHST, however, there is uncertainty (ie, role of chance = 5%) associated with this estimation. This implies that if samples were taken repeatedly, 95 times out of 100, the range would contain the population effect. The width of the confidence interval is determined based on the confidence a researcher must have in this decision-making process. Higher levels of confidence (99% CI vs 95% CI) produce wider confidence intervals. In addition to the level of confidence, sample size and the variability in the data also have an impact on the width of the confidence interval. As discussed previously, the *P* value only indicates statistical significance. In contrast, confidence interval can assess both statistical and clinical significance. If a 95% CI for the mean difference contains zero, there is no difference between the 2 groups. Contrary, if the 95% CI does not contain zero, there is a statistically significant difference. **Table 5** provides 3 examples of research questions, the statistical techniques best suited for answering the questions, and the reporting of statistical findings, including effect sizes, and confidence intervals.

UNDERSTANDING MINIMAL DETECTABLE DIFFERENCE AND MINIMAL IMPORTANT DIFFERENCE

Although understanding reliability, the extent to which a measure is free of error, may seem simplistic, it is a challenging concept to decipher. There are 2 types of errors: systematic and random error. These added together are the total measurement error associated with a measure. Systematic errors could occur due to factors, such as learning effects or participant fatigue across multiple trials, whereas random errors take place due to biological, mechanical, or research protocol changes. There are 3 frequently used statistical methods to report reliability: relative reliability (ie, test-retest correlation), systematic bias (ie, change in mean), and absolute reliability (ie, repeated measurement variability).[10,11] Relative reliability is the consistency of an individual's position within a group over several measurements. The intraclass correlation coefficient is commonly seen within sports medicine literature as a measure of relative reliability (ranging from 0 to 1, with 1 perfect relative reliability). Systematic bias can be reported using a paired *t* test (2 measurements) or a repeated measures analysis of variance (3 or more measurements). Although relative reliability and systematic bias are important, absolute reliability may have more clinical relevance. Absolute reliability indicates variability in repeated measurements for an individual. Standard error of measurement (SEM) is 1 method of determining absolute reliability. The SEM is typically seen when the intraclass correlation coefficient is presented. In a similar manner, the MDD also provides measurement error boundaries; however, it is more conservative (wider error boundaries) than the SEM. Any change that occurs outside the MDD boundaries can be considered more appropriately as real change. The MID is the level of change, beyond the MDD, that a patient perceives to be meaningful and at which the patient would repeat the treatment again.[12] Two approaches are used to determine the MID, anchor based (an external criterion is used to determine if a meaningful change has occurred or not) and distribution based (statistical significant change, precision of the measurement, or variability in the scores). The second approach is more statistical in nature and does not account for patient preference as much as the anchor-based methods. The anchor-based methods, however, lack the generalizability of the distribution-based approach. There is little agreement in

Table 5
Statistical approaches frequently seen in literature with the common assumptions, typical reporting of statistical findings, and clinical interpretation to assist a practitioner reading a research article

Research Question	Analytical Technique	Common Assumptions	Statistical Findings	Clinical Interpretation
Is there an association between injury location (knee or hip) and abnormal LSI (asymmetry $\geq$15% or asymmetry <15%)?	Chi-square test of association	Observations are independent; expected frequency is at least 5 per cell	$\chi^2_1 = 0.400$, $P = .527$, $R = 0.10$, 95% CI (−0.208, 0.390)	There is no association between the location of injury and abnormal LSI results.
Is there a difference in LSI depending on type of sport (soccer, basketball, or football)?	One-way ANOVA with Tukey post hoc testing	Observations are independent; homogeneity of variance; populations follow a normal distribution	$F_{2,37} = 1.674$, $P = .015$, 95% CI: (87.311, 91.870), $\eta^2_p = 0.20$; football to basketball: $P = .035$	There is a difference in LSI based on the type of sport, specifically between football and basketball athletes.
Do gender, age, and LSI predict time to RTP after a quadriceps strain?	Multiple linear regression	Observations are independent; homogeneity of variance; normal distribution; linearity; noncollinearity;	$F_{3,36} = 2.990$, $P = .044$, $R^2 = 0.20$; LSI: $t = 2.241$, $P = .031$, 95% CI: (0.011, 0.0229)	The overall model (gender, age, and LSI) can predict days until RTP in those athletes with a quadriceps strain; specifically, LSI is statistically significant.

When statistical results are reported, it is customary to report the value of the test statistic (ie, F, *t*, and χ^2) and the degrees of freedom as a subscript. Effect sizes (ie, *d*, *R*, η^2_p, and R^2), and confidence intervals (upper and lower boundaries separated by a comma) appear after the *P* values.

Abbreviations: ANOVA, analysis of variance; R^2, coefficient of determination; RTP, return to play.

literature on which approach is superior.[13–16] A practical example of using the MDD and MID for interpreting a patients' change in shoulder pain has recently been provided.[17]

INTERPRETING STUDIES THAT DO NOT REACH STATISTICAL SIGNIFICANCE

Currently, there exists a publication bias, with largely only articles yielding statistically significant positive results published.[18–20] The bias is likely a combination of researchers hesitant to submit research with nonsignificant results and journal editorial boards tending to accept articles only with statistically significant results. Regardless of the source, the outcome presents a challenge when conducting systematic reviews and meta-analyses because only studies with statistical significance, and likely large effect sizes, are available for consideration. For example, in the context of sports medicine research examining various interventions for injury prevention and recovery, this may lead to an overestimation of the benefits various interventions may have on preventing injury and improving patient function. Thus, it is important that high-quality studies with nonsignificant findings be published; the challenge is how to evaluate whether such studies meet the bar for high quality.

Statistical power is the probability of attaining statistical significance ($P<\alpha$) and not committing a type II statistical error. In other words, statistical power is the likelihood of rejecting a false null hypothesis and making an appropriate statistical decision. Statistical power is influenced by the α level, sample size, and magnitude of the treatment effect relative to the variation in how the participants responded to the treatments (ie, effect size). As discussed previously, an α level of .05 is standard practice. In cases of intervention studies, the magnitude of the treatment effect and response variation is largely a function of the intervention parameters selected, such as dosage (eg, how many rehabilitation sessions per week), as well as the participant inclusion and exclusion criteria (eg, severity and variability in pathology). It is easier to achieve statistical significance when there are large effects and little variation in the response to the treatment. Sample size is the power parameter that often receives the most attention during the study planning process. Prudent practice is to conduct an a priori power analysis to estimate how many participants are needed to reach statistical significance given chosen values for the α level, statistical power (.80 is most common), and the anticipated effect size. Preferably, the anticipated effect size should be based on what is considered a clinically relevant change. Using the MDD or MID can help with selecting an effect size that is clinically relevant. Previous literature and pilot studies can also provide guidance regarding the potential potency of an intervention. Although an a priori power analysis can be conducted by using the standard interpretation conventions, described previously (eg, Cohen's *d* classifications), the authors suggest avoiding this approach in isolation because it does not consider the clinical relevance of the change.

Readers of research should look for evidence of an a priori power analysis, particularly in studies reporting nonsignificant results. In the circumstance of nonstatistically significant results, evaluation needs to occur to determine which of the following is more tenable: (1) Is there truly no difference in the effectiveness of the treatment interventions? or (2) Is there a possibility that type II statistical error has occurred? Along with examining other elements of the research design and execution, the details of an a priori power analysis can help with this evaluation. If an a priori power analysis is not included or if insufficient detail regarding the statistical power analysis is provided, determining which of the 2 scenarios, discussed previously, is more likely cannot occur. **Box 1** contains some guidance for evaluating an a priori analysis.

Box 1
Some considerations for evaluating an a priori statistical power analysis

Were the chosen alpha and power levels appropriate?

Is there sufficient detail regarding the source of the difference and variance estimates to evaluate? If yes:
- Do the difference/variance (ie, effect size) have clinical relevance?
- If previous research/pilot studies were used to assist, was the population similar to the current study?
- Was there some buffer to account for sampling fluctuation included?

In cases of an intervention study failing to reach statistical significance, assuming the research was well designed and executed, answering "yes" to each of the questions suggests there being truly no effect or a smaller than anticipated effect.

Finally, even with an optimally conducted a priori power analysis to establish sample size prior to commencing the research, it is still possible that non–statistically significant results could be due to a type II statistical error. Again, research relies on sampling from populations and, therefore, the study sample may demonstrate higher variability than the power analysis estimate or illicit less response to the intervention than the population. Thus, similar to needing study replication when statistically significant results occur to provide more evidence against a possible type 1 error, well-conducted investigations with solid rationale that fail to reach statistical significance may also warrant independent replication to rule out a possible type II error.

SUMMARY

Few practitioners enter sports medicine with a desire to develop an extensive understanding of statistics; however, as part of the EBP movement, practitioners must have some competency in understanding research. Interpreting the statistical analysis and results is an important component of that competency. This article provides some of the minimal statistical essentials sports medicine practitioners can use to facilitate understanding of research reports. Statistical significance, as reflected by a *P* value associated with a statistical test, indicates the influence of chance on the outcome, whereas clinical meaningfulness reflects the degree to which the study results reported are relevant to sports medicine practice. When statistically significant differences are revealed, confidence intervals and effect sizes can be used to enhance the practical interpretation of the research results. Absolute reliability characteristics, such as the MDD, determine the extent of error around a measurement and, when coupled with an appropriate MID estimate, can assist in triangulating clinically meaningful changes in patients undergoing treatment. In the circumstance of nonstatistically significant results, evaluation needs to occur to determine if the study had adequate statistical power prior to concluding the data are indicating that no difference or association exists in the population. Better understanding of these research and statistical concepts will improve EBP and, therefore, patient care.

REFERENCES

1. Goodman S. Toward evidenced-based medical statistics. 1: the *P* value fallacy. Ann Intern Med 1999;130:995–1004.
2. Kline R. Beyond significance testing: statistical reform in the behaviorial sciences. 2nd edition. Washington, DC: American Psychological Association; 2013.

3. Blume J, Peipert J. What your statistician never told you about P-values. J Am Assoc Gynecol Laparosc 2003;10(4):439–44.
4. Huck S. Reading statistics and research. Boston: Pearson Education, Inc; 2012.
5. Riemann BL, Lininger M. Statistical primer for athletic trainers: the difference between statistical and clinical meaningfulness. J Athl Train 2015;50(12):1223–5.
6. Kirk ER. The importance of effect magnitude. In: Davis S, editor. Handbook of research methods in experimental psychology. Oxford (United Kingdom): Blackwell; 2003. p. 83–105.
7. Lininger M, Riemann BL. Statistical primer for athletic trainers: using confidence intervals and effect sizes to evaluate clinical meaningfulness. J Athl Train 2016; 51(12):1045–8.
8. Cohen J. Statistical power analysis for the behavioral sciences. 2nd edition. Hillsdale (NJ): Lawrence Erlbaum Associations, Publishers; 1988.
9. 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): 918–20.
10. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998;26(4):217–38.
11. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc 1999;31(3):472–85.
12. Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007; 7(5):541–6.
13. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56(5):395–407.
14. de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied “minimally important change” values. J Clin Epidemiol 2010;63(1): 37–45.
15. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2011;11(2):171–84.
16. Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. COPD 2005;2(1):57–62.
17. Riemann B, Lininger M. Statistical primer for athletic trainers: the essentials of understanding measures of reliability and minimal important change. J Athl Train 2018;53(1):98–103.
18. Hopewell S, Loudon K, Clarke MJ, et al. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009;(1). MR000006.
19. Joober R, Schmitz N, Annable L, et al. Publication bias: what are the challenges and can they be overcome? J Psychiatry Neurosci 2012;37(3):149–52.
20. Torgerson C. Publication bias: the Achilles’ heel of systematic reviews? Br J Educ Stud 2006;54(1):89–102.