GETTING THE MOST OUT OF RESEARCH

Looking beyond the *P*-value

J. A. LaCross

Department of Obstetrics and Gynecology, Michigan Medicine, Ann Arbor, and Program in Physical Therapy, University of Michigan–Flint, Flint, Michigan, USA

Abstract

Many physiotherapists find it challenging to apply research findings in practice because there is uncertainty about statistical interpretation. This commentary aims to explain three statistics that are often seen in the results sections of research papers, i.e. probability values, confidence intervals and effect sizes. Additionally, the importance of considering these values together is discussed. A better understanding of statistical interpretation could help physiotherapists to use research more effectively to guide care.

Keywords: confidence interval, effect size, results, statistics, translation.

Cite as: LaCross J. A (2024) Looking beyond the *P*-value. *Journal of Pelvic, Obstetric and Gynaecological Physiotherapy* **134** (Spring)**,** 67–72.

DOI: 10.62399/QFNC5168

Introduction

Do you ever find yourself skimming over or completely skipping the statistics reported in the results section of a paper? If so, do you ask yourself why? This frequently occurs because reading something that is difficult to understand is frustrating and occupies time that busy clinicians do not have. Physiotherapists go to the literature with a specific purpose: to find out how to better assist their patients.

The aim of the present commentary, which is a follow-up to LaCross (2023), is to explain three common statistics, probability values (*P*-values), confidence intervals (CIs) and effect sizes, and how to interpret this information. Each statistic is discussed in turn, and the benefits of considering these together is then discussed. A summary of what each statistic is and is not is provided at the end of each section. Finally, an example is given to showcase all the concepts reviewed. Learning how to interpret these statistics together will improve clinicians' ability to translate research findings into practice more effectively.

Correspondence: Jenny LaCross PT DPT PhD ATC CLT, Department of Obstetrics and Gynecology, Michigan Medicine, 1500 East Medical Center Drive, Ann Arbor, MI 48109, USA (e-mail: jalacros@umich.edu).

Summary:

- *What this commentary is:* a simple overview of three common statistics, and how to use these to interpret results more effectively.
- *What this commentary is not:* a comprehensive overview aimed at developing clinician-researchers.

How do statistics help busy clinician readers?

Physiotherapists use research literature to answer clinical inquiries. These can be background questions (e.g. What is the underlying pathophysiology of endometriosis?), or foreground ones (e.g. In patients with endometriosis, is myofascial release more effective than aerobic exercise for pain reduction?). Background questions ask for general information about a condition or disease process. Foreground questions ask for specific data to better inform clinical decision-making.

When asking a foreground question, clinicians are interested in an answer that will help them to improve the care that they give the patients in front of them. Because the patients have probably not been studied, clinicians look instead for papers investigating groups of individuals with the same profile (e.g. females with endometriosis). The goal is to apply the information gathered

Table 1. Example of hypothesis development: (research question) What is the effect of soft-tissue mobilization on self-efficacy scores in patients with endometriosis?

Hypothesis type	Example	Meaning
Null	Soft-tissue mobilization will *have no effect* on pain scores in patients with endometriosis	Scores will not change from pre- to post-treatment
Alternative non-directional	Soft-tissue mobilization will *have an effect* on pain scores in patients with endometriosis	Scores will change from pre- to post-treatment
Alternative directional	Soft-tissue mobilization will *decrease* pain scores in patients with endometriosis	Scores will decrease, or clinically improve, from pre- to post-treatment
Alternative directional	Soft-tissue mobilization will *increase* pain scores in patients with endometriosis	Scores will increase, or clinically worsen, from pre- to post-treatment

from this group, or sample, to a specific patient. However, it is important to remember that statistics do not provide absolute answers to clinical questions about a single patient. Individual studies sample the population of interest to obtain a group of individuals with the hope that the sample selected represents the population at large. Samples are used in studies because it is impossible to include the entire population of interest (e.g. every female with endometriosis). The distinction between a sample and a population is important because many statistics provide information about how representative the study sample is of the population.

Statistics is a mathematical science that quantitatively summarizes information for readers to interpret. The relevant data are generally, although not exclusively, found in the results section of a research article, and can be represented within the text as symbols and numbers, or in images such as graphs. Three statistics are commonly reported to help the reader interpret the information collected in a study: *P*-values, CIs and effect sizes.

Summary:

- *What statistics does:* uses mathematical modelling informed by data collected from a sample to predict outcomes and/or detect differences in a population of interest.
- *What statistics does not do:* provide absolute answers to clinical questions about a single patient.

P-values

Probability values are perhaps the most-reported and well-recognized form of statistics. A *P*-value is the long-run probability of how likely it would be to obtain a value of a test statistic at least as big as the one found if the null hypothesis is true (Field 2018). However, the *P*-value is affected by the sample size: samples that are too large or too small negatively affect the accuracy of the information provided. For additional information on sample size, refer to Kamper (2022). To provide additional context for the *P*-value, null hypothesis significance testing (NHST) must be reviewed.

Null hypothesis significance testing is the basis for most interventional rehabilitation research. It is used to answer questions like, “What is the effect of intervention X on symptom Y?” To answer this statistically, there must be a null hypothesis, and an alternative or research hypothesis. The null hypothesis is written to reflect that a given intervention results in no effect or change. An alternative hypothesis states that a given intervention does result in an effect or change, and can be written in two ways. First, the alternative hypothesis can be non-directional, meaning that an effect or change will occur, but the direction of that effect is not stated. It could be positive or negative. Secondly, an alternative hypothesis can also be directional, meaning that an effect or change will occur, and the direction of that effect is specified. Table 1 provides an example. The way that the alternative hypothesis is written affects how the *P*-value is interpreted.

Before a study is carried out, or *a priori*, an α value must be set. This is the predetermined significance level of how frequently the researcher is prepared to be wrong. This type I error rate is the probability of accepting an effect in a population as true when no such effect exists (Field 2018). The type I error rate for a given study is either 0 or 1, i.e. an error of this kind was either made or not, but there is no way to know. This is why the α set *a priori* is the long-term error rate. This threshold of significance is usually set at 0.05 or 5%, but it could be a different value depending on what is appropriate for the study in question. An α of 0.05 means that, when there is a 5% chance (or 0.05 probability) of getting the result obtained if no effect exists and the null hypothesis is true, the researchers can

Table 2. To accept or reject the null hypothesis

Condition	How to interpret the null hypothesis
If *P*-value is ≤pre-determined α	Reject
If *P*-value is >pre-determined α	Accept

be confident enough to accept the effect as real and not occurring by chance (Field 2018). The *P*-value from the study sample is then compared to this α value. For information on interpreting this comparison, see Table 2.

The other type of mistake that can be made is a type II error, or accepting the null hypothesis when it is false. This is rejecting a true effect that exists. It tends to occur when the sample size is too small, which is why the *P*-value must be considered within the context of the sample size. *A priori* power calculations, which determine the number of participants needed to find an expected difference, are commonly reported in the methods section of a research paper. Statistical power is the probability of either avoiding a type II error, or rejecting the null hypothesis when the alternative is true (Field 2018). If it is reported, this calculation can be referred to when reading the results to ensure that the number of participants included meets the predetermined threshold. If not, the study may be underpowered, which is sometimes listed as a study limitation in the discussion section of a research paper.

Finally, the difference between statistical significance, as indicated by the *P*-value, and clinical relevance must be considered. A *P*-value only indicates if an effect exists (assuming that the sample size is large enough), not how big or meaningful it is. In large samples, *P*-values can be significant when the effect is very small. Conversely, small samples may yield non-significant *P*-values, but meaningful effects. When using NHST, a *P*-value that exceeds the pre-determined α value, i.e. $P > 0.05$, indicates that an effect was not big enough to be found based on the sample size used, not that the effect is zero. Again, this is why the *P*-value must be interpreted in the context of the sample size and effect size.

Summary:

- *What a* P-*value is:* the long-run probability of how likely it would be to obtain a value of a test statistic at least as big as the one found if the null hypothesis, i.e. that the intervention results in no effect or change, is true.
- *What a* P-*value is not:* the probability (1) of a chance result, (2) that the null hypothesis is true or (3) that the alternative hypothesis is true.

Confidence intervals

Confidence intervals are another commonly reported statistic. These provide additional information about a parameter, such as the sample mean. A CI is the interval, or limits, that contain the true population value of the parameter in a certain percentage of samples (Field 2018). This percentage is usually 95%, but can also be 99%. This is why it is called a 95% CI. This interval is written in parentheses following a parameter value, "(95% CI: lower limit–upper limit)", or is visually represented as error bars on a graph.

Confidence intervals are important because a parameter (e.g. an average or mean reported from a sample) is not the true population value. These are boundaries within which the population mean is believed to fall based on the sample mean (Field 2018). If a study was performed using a sample of 30 people with limited hip flexion range of motion (ROM), and following a self-mobilization exercise, flexion improved to a mean of 120°, this indicates that the mean post-treatment hip flexion ROM for this sample is 120°. It is not the mean for the population, which would consist of everyone with limited hip motion who performed the self-mobilization exercise. As noted earlier, it is impossible to test the entire population, and therefore, the population mean is estimated using this sample mean. By reporting the sample mean and the 95% CI, we can see the range of values within which the population mean would fall in 95% of samples. If the 95% CI from the example was 117–125, then the population mean ROM would fall between 117° and 125° in 95% of samples. In 5% of samples, the mean would fall outside of these limits, either below 117° or above 125°.

In this example, the 95% CI is narrow, which indicates that there is less variability in the sample measure. The sample mean is probably close to the population mean, indicating that the sample selected is likely to be representative of the population of interest. Wide CIs suggest that there is more variability in the sample measure. The sample mean may be quite different from the population mean, indicating that the sample may be a poor representation of the population. This is important for clinicians to consider. Graphical representations of CIs are show in Figure 1.

Returning to the example used in Table 1, an investigation of the effect of soft-tissue

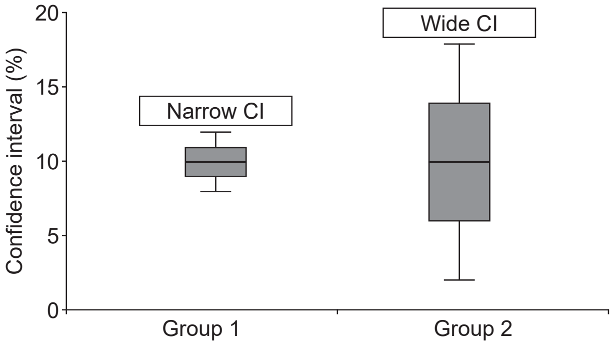

Figure 1. Visual representations of two different 95% confidence intervals (CIs): (group 1) a narrow CI; and (group 2) a wide CI. N.B. Both groups have the same mean and median, which is represented by the dark black line in the centre of each box. Error bars represent the 95% CI of the mean.

mobilization on pain scores in patients with endometriosis, suppose that this study included 50 females with endometriosis whose average pain score on the Numeric Pain Rating Scale following the intervention was 2. Now, assume that the 95% CI of the post-intervention pain score is 1–7). What does this indicate? First, the reader needs to decide whether this is a narrow or wide interval, given the measure used. The Numeric Pain Rating Scale ranges from 0 to 10, so a 95% CI of 1–7 is wide. Next, the CI indicates that, in 95% of samples, the population value of average post-intervention pain could fall anywhere between 1 and 7. It is possible that the 50 participants included in this study sample are not representative of the population of all females with endometriosis. If this is the case, then clinicians may need to consider whether soft-tissue mobilization is a good intervention for their patient with endometriosis.

Another situation to pay close attention to is when a 95% CI for a change score or mean difference includes 0. This occurs when the lower boundary is negative and the upper one is positive; for example, 95% CI: –5–3. Confidence intervals that include 0 indicate the possibility that there is no effect or no difference between groups. In the context of NHST, the null hypothesis would not be rejected if the CI includes 0. For studies examining the effectiveness of an intervention, if the intervention demonstrates an improvement in the outcome of interest (e.g. pain or ROM) compared to a control group, but the 95% CI associated with the mean improvement includes 0, it is possible that this result could be caused by chance and the intervention itself did not change anything. For additional information on CIs, refer to Kamper (2019).

Summary:

- *What a 95% confidence interval is:* the interval, or limits, that contains the “true” population value of the parameter in 95% of samples.
- *What a 95% confidence interval is not:* a 95% chance that the interval contains the population value, i.e. subjective 95% confidence that the value of the population parameter will fall within the given interval.

Effect sizes

Effect sizes are especially important statistics to consider when assessing the clinical significance of a result. An effect size is a standardized measure that describes the size or magnitude of an effect (Field 2018). It helps to provide context for interpreting the *P*-value. Unlike a *P*-value, effect size is not affected by sample size. In general, smaller values indicate smaller effects and *vice versa*. When interpreting effect sizes, it is important to remember that small intervention effects may still be very clinically important. Commonly reported effect sizes are Cohen’s d and Pearson’s r.

Cohen’s d is used to compare group means. Pearson’s r assesses the linear relationship between two variables. Researchers indicate specific value ranges associated with small, medium and large effect sizes for a study in the methods section of a research paper because these vary. For example, effect sizes are considered small if $d = 0.2–0.5$, medium if $d = 0.51–0.8$ and large if $d > 0.8$. Effect sizes can also provide information about the direction of a relationship. A positive Pearson’s r indicates that two variables move in the same direction, i.e. as one increases, so does the other. A negative r shows that two variables are inversely related, i.e. as one increases, the other decreases. Readers should consider effect size in the context of the *P*-value and confidence interval when translating research into practice.

Summary:

- *What an effect size is:* a standardized measure that describes the size or magnitude, and sometimes direction, of an effect.
- *What an effect size is not:* a test of significance or a statistic based on sample size.

Putting it all together

It should now be clear what *P*-values, CIs and effect sizes are, and what kind of information each provides. To summarize:

- *P*-values tell the reader whether an effect probably exists or not on the basis of the sample;

Table 3. Summary considerations for the clinician reader

Statistic	Considerations
P-value	Was the sample size large enough? Did the number of participants meet the number determined in the *a priori* power calculation?
Confidence interval	Is it narrow or wide? Does it include zero?
Effect size	Is this reported? (Sometimes it is not) Is the size of the effect clinically meaningful?

- 95% CIs provide the range into which the population parameter would fall in 95% of samples; and
- the effect size reveals the size or magnitude of the effect independent of sample size.

For additional summary considerations, see Table 3.

Interpreting these statistics together provides a better picture of what the results indicate, and helps clinicians to apply research in practice effectively.

To conclude, a fictional rehabilitation example of interpreting these statistics together is provided below. It highlights a circumstance in which only considering the *P*-value would be misleading. Next time you pick up a research paper to help answer a clinical question, please look beyond the *P*-value.

Example

Disclaimer: This is fictitious and designed for learning purposes only.

Study title: "ABC training is more effective than pelvic floor muscle exercises for reducing urinary leakage".

Study background. Researchers are studying the effects of a new physical therapy intervention, i.e. ABC training, to reduce urinary leakage compared to a standard care regimen of pelvic floor muscle exercises performed for 8 weeks. The α was set at 0.05. A urinary incontinence questionnaire was used to measure subjective urinary leakage at the beginning and end of the 8-week intervention. This questionnaire uses an 11-point scale: (10) severe leakage; and (0) no leakage.

Results. After conducting a large-scale randomized clinical trial with 10 000 participants, the researchers found the following:

- mean urinary leakage reduction in the ABC group = 1.2;
- mean urinary leakage reduction in the standard care group = 1;
- mean group difference = 0.2 (95% CI: –0.1–2.5), $P < 0.001$; and
- calculated effect size, $d = 0.2$.

Interpretation. The *P*-value of < 0.001 suggests that the difference between the ABC training and standard care groups is highly statistically significant at the conventional α level of 0.05. Small effects can be statistically significant in large samples. The titles of papers, such as the one in this example, can be misleading if based only on statistical significance.

Confidence interval: Given that the 95% CI includes 0, the null hypothesis that there is no difference between groups cannot be excluded.

Effect size: The Cohen's *d* value of 0.2 indicates that the magnitude of this difference is small. The clinical significance of a 0.2-point difference on an 11-point pain scale may be negligible.

Conclusion: While the results show that ABC leads to a statistically significant reduction in urinary leakage compared to standard care in a very large sample, the magnitude of this difference is probably too small to be clinically meaningful. Additionally, the 95% CI includes zero, suggesting there may be no difference between the two groups. Clinicians should weigh these findings against other considerations, such as costs or potential side effects when deciding whether or not to implement ABC in their practice. A more-appropriate title for this study would be "ABC training is no more effective than pelvic floor muscle exercises for reducing urinary leakage".

Acknowledgements

The author would like to thank Cara Lewis PT PhD for her invaluable assistance with the refinement of the concept of this commentary, and also her technical support with it.

Ethical approval

Ethical approval from the institutional review board was not obtained because this commentary did not include the use of human subjects.

Funding

This project was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (grant number U2C DK129445-02). It played no role in the writing of the paper or the decision to submit it for publication.

Conflicts of interest

The present author does not have any conflicts of interest to declare.

References

Field A. (2018) *Discovering Statistics Using IBM SPSS Statistics*, 5th edn. SAGE, Thousand Oaks, CA.

Kamper S. J. (2019) Confidence intervals: linking evidence to practice. *Journal of Orthopaedic & Sports Physical Therapy* **49** (10), 763–764.

Kamper S. J. (2022) Sample size: linking evidence to practice. *Journal of Orthopaedic & Sports Physical Therapy* **52** (8), 563–564.

LaCross J. A. (2023) Why understanding study methods matters. *Journal of Pelvic, Obstetric and Gynaecological Physiotherapy* **133** (Autumn), 47–52.

Jenny LaCross PT DPT PhD ATC CLT (she/her) is a board-certified clinical specialist in women's health physical therapy, and a postdoctoral research fellow with the Pelvic Floor Research Group at the University of Michigan–Ann Arbor. She also works as adjunct faculty for South College and University of Michigan–Flint in these institutions' respective clinical doctorate of physical therapy programmes. As a clinician scientist, Jenny understands the difficulty of translating research into practice, and aims to help other clinicians bridge this gap and improve the delivery of care.