SPECIAL COMMUNICATION

Statistical and Clinical Interpretation of Research Results

Garry T. Allison, PhD*

There is a well-known phenomenon of publication bias toward manuscripts that report statistically significant differences. The clinical implications of these statistically significant differences are not always clear because the magnitude of the changes may be clinically meaningless. This article relates the critical *P* value threshold to the magnitude of the actual observed change and provides a rationale for reporting confidence intervals in clinical studies. Strategies for improving statistical power and reducing the magnitude of the confidence interval range for clinical trials are also described. (J Am Podiatr Med Assoc 97(2): 165-170, 2007)

In research, the purpose of conducting statistical tests is to determine the likelihood of the observed changes occurring over and above the probability of the event happening by chance. With this information, readers can conclude with a level of confidence that the findings were due to a nonrandom event, eg, the change in rearfoot angle associated with the fitting of an orthosis. The difficulty in interpreting clinical research is that statistically significant findings may not be clinically useful.

Publication Bias Is Reflected in the Research Process

A well-designed research project will incorporate sufficient methodologic rigor to allow confidence in validly testing the proposed hypothesis. The confidence is related to the *P* value, and this has a great effect on the potential for the paper to be published because readers, authors, and reviewers tend to have a bias toward manuscripts that show statistically significant results. Publication bias in scientific journals is the greater likelihood of publication of manuscripts with statistically significant results.[1-3] Furthermore, evidence suggests that authors of papers reporting clinical trials with nonsignificant results tend to take longer to submit a manuscript for review, if they submit it at all.[1, 4, 5] Therefore, there is pressure on researchers to reach a statistically significant result. In podiatric medicine, this pressure may be seen in the conundrum of analyzing feet or sides separately or pooling the data to increase the sample size and, therefore, the chance of detecting a statistically significant *P* value.[6-8]

Although the decision to pool or not to pool feet relates to the statistical validity and the research question being asked,[9] there are other aspects of data interpretation that may be considered over and above using the *P* value threshold. The interpretation of statistical power, the magnitude and distribution of the change in the clinical measurement, and, finally, the interpretation of these reported changes in the clinical context of what is or is not trivial are key interpretive factors that the podiatric physician (reader and researcher) must consider.

The *P* Value

The *P* value is a number between 0 and 1 that indicates the probability of the result happening by chance. It is standard scientific practice to set a threshold of statistical confidence by defining the *P* value at which findings will be deemed "statistically significant." Essentially, the α level or *P* value threshold is the proverbial line drawn in the sand where the reader can state that at this level of confidence it can be asserted that there is a statistically significant difference. So, in studies in which a 95% probability limit is set (ie, $P < .05$ will be considered statistically significant), there is a 5% chance that the result occurred by chance. Note that the distinction between two different *P* values is one of confidence that a statistically significant result was not due to random chance. It does not provide readers with an indication of the magnitude of the actual change.

In the context of interpreting the impact of orthoses, clinicians should look at the *P* value, which

*School of Surgery and Pathology, Centre for Musculoskeletal Studies, Level 2 MRF Bldg, Rear 50 Murray St, Perth, Western Australia 6014, Australia.

indicates the probability of the event happening by chance. Values of $P = .040$ and $P = .060$ provide 96% and 94% confidence, respectively, that the results did not happen by chance. Each finding falls on either side of the magical $P = .050$ or 95% threshold for statistical significance. The selection of the 95% threshold is a scientific standard; however, if the threshold was set at 90%, the confidence limits would become tighter, and any P value less than .10 would reach statistical significance. For this hypothetical analysis (Table 1 and Figure 1 derived from raw data[7]), one could see that at the 90% confidence interval ($P = .100$), the left foot would not detect a significant difference ($P = .103$), yet the right foot would ($P = .077$). Individuals with complete devotion to the P value in this example would then be searching for ("sinister") reasons why orthoses work better on the right foot.

While balancing these P value threshold issues, the reader has to look at the actual magnitude of the changes in terms of clinical significance. These interpretations are important in clinical decision making and should influence the discussion of research manuscripts that reflect clinical importance as much as the P value. Pooling the data where it is applicable to do so just narrows the confidence limits; the P value reaches the 95% threshold, but in reality the mean and range of the treatment effect, in this example, did not change substantially.

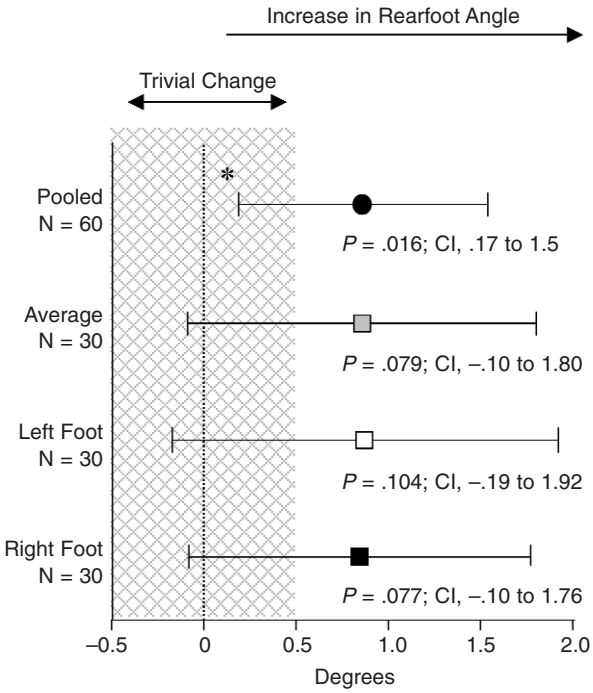

Figure 1. Mean difference in the rearfoot angle with and without an orthosis for unilateral, average, and pooled data. The hypothetical data set is from Menz.[7] Shaded area may reflect detected differences that are smaller than the measurement error or the clinical significance defined by the clinician and, therefore, may be considered trivial. Error bars represent 95% confidence intervals (CIs). *Statistically significant difference ($P < .05$).

Insufficient Power

There is often a gap between clinical data collection and ideal research paradigms, and this reflects the importance of having sufficient power to detect a difference. Equally critical is the understanding that if a suitable statistical test does not detect a difference, then the study is underpowered to do so. Undertaking *post hoc* power analysis to show that a nonsignificant test was underpowered is at best wasted effort because the P value and the power of the study are related.

Low power, such as a nonsignificant P value, is evidence that the difference between the conditions was small relative to the variance in the data. Therefore, in this specific case the statistical test did not reject the null hypothesis, ie, that there is a difference. The basic principle of statistics is that we can only show a level of confidence that things are different, so if no difference was detected, it is not necessarily

Table 1. Paired *t* Test Results for Rearfoot Angle With and Without an Orthosis for Unilateral, Mean, and Pooled Data Sets

Data Set	Feet (No.)	Orthosis Condition		*P* Value
		With	Without	
Right foot only	30	4.27 (1.87)	3.43 (1.78)	.077[a]
Left foot only	30	3.87 (2.06)	3.00 (1.62)	.104
Mean (right and left)	30	4.07 (1.96)	3.22 (1.58)	.079[a]
Pooled feet	60	4.07 (1.96)	3.22 (1.70)	.016[b]

Note: Data are given as mean (SD). Hypothetical data are from Menz.[7]
[a]Significant at $P = .10$.
[b]Significant at $P = .05$.

correct to say that they are the same. This is known as a "type II statistical error." Readers should not fall into the assumption that nonsignificant difference implies that the results are the same. Similarly, there are many complex statistical arguments that if the hypothesis is rejected then there are a multitude of alternative hypotheses that could be generated, and, therefore, there is no additional statistical proof that one particular alternative hypothesis should be accepted.[10-12]

By reporting that a study is low in statistical power the author may define strategies to improve the power of the study and acknowledges weaknesses in the study design. The statistical power of a study can be improved using many strategies that essentially are related to a process of 1) increasing the treatment effect and 2) reducing the estimated variance in the data. Table 2 illustrates examples from each of these categories. By pooling feet data we see a doubling of the sample size, and the estimates of the spread of the data become more precise; therefore, statistically significant differences are more likely to be detected. Figure 1 shows that the confidence limits become smaller when the sample size is doubled (pooled feet data). The magnitude of these changes, reflected in the mean and confidence limits of the change scores, show minimal changes in the estimation of the clinical effect in any combination of the single or pooled data. The mean impact of orthoses (from this hypothetical data) is approximately 0.85° for the left (0.87°) and right (0.83°) feet, and the range of the effect with a 95% level of confidence is between –0.2° and 1.9°.

A clinician should not be swayed by the statistical significance (*P* value) because the magnitude of the changes for statistically significant and statistically nonsignificant differences are similar. With the reporting of confidence limits, the clinician should be able to consider what proportion of the range of the confidence interval lies above the threshold of clinical significance.

The magnitude of what can be expected to be meaningful depends on the type of research being conducted and the individual interpretation of the actual data. By reporting change scores and 95% confidence limits, the authors provide the information within the manuscript to allow the different readers to interpret the findings. When *P* values are the only indicator of "significance," interpretation by the reader is limited to statistical confidence, which may have limited value in the clinical setting.

Table 2. Strategies to Improve the Statistical Power of a Study by Increasing the Effect Size and Decreasing Variance

Strategy	Examples
To increase the effect size	
Target specific populations with known impairments	Select individuals who have a pathologic rearfoot position rather than a normal rearfoot position
Use outcome measures that are specific and sensitive to the mechanism/action of the intervention	Specific scales focusing on gait biomechanics may be more sensitive to testing mechanisms Alternatively, use a patient-specific scale related to activities that are limited by foot pain for treatment efficacy studies
Provide optimal treatment	Ensure that the full treatment is provided and that the intervention is matched to the individual; have a run-in period before randomization to minimize dropouts
Control for known covariants that may inhibit/diminish the response to treatment	Target populations that have the greatest chance of responding to the treatment and limit adjunct treatments being performed in the control group
To decrease population variability	
Use measurements with a low level of random error	Know the measurement error of your instruments and develop methods to minimize variance; time taken here can result in fewer subjects required
Select a homogeneous population	Be strict on inclusion criteria to narrow the variability in the test population; this is balanced with the generalizability of the findings
Increase the sample size	Undertake a pilot study to determine estimated sample variance, treatment effect size, and, therefore, numbers needed; if it is too many, change methods
Control for extraneous factors that may increase variability in outcome measures	Identify and measure prognostic factors that may affect responsiveness to treatment; use these as a covariant in the final statistical analysis (analysis of variance)

How Does the Reader Define a Meaningful Change?

The magnitude of a meaningful change depends on the type of research question and the context of the study. There are various terms in the literature, and each represents a different threshold at which a reader may consider that a significantly meaningful change has occurred (Table 3).

Generally, there is agreement that for a measurement to be meaningful in any context it must be reliable and valid. In terms of reliability, the general rule is that the threshold of what is considered to be meaningful change must be larger than the measurement error, noise, or typical error.[13, 14] Otherwise, the real changes are difficult to observe.

Three sources of measurement error are error or variation from the instrument, the individuals conducting the test, and the subjects. Good reliability studies can identify, with some confidence, the level of typical error attributable to combinations of one or more of the three sources of error. For clinical interpretation, there is an advantage to reporting the magnitude (mean and confidence limits) of the error in the units of the measurement that the clinician understands. This at least provides a benchmark to determine whether the level of error is comparable with what the clinician may personally consider clinically important. The reporting of *P* values and population-dependent correlations (eg, intraclass correlations) provides estimations of the measurement error in a proportional comparison with the variance in the population tested. The use of these correlation coefficients, even if confidence limits are provided, shows less clinical utility to the reader. Inherently, all reliability error estimates depend on the generalizability of the results. The ability to generalize reliability error estimates reported in the research literature depends on many factors, including the person performing the test, the equipment, and the population being tested.

Once the typical error associated with an instrument, tester, and study population is determined, it is possible to determine the probability of detecting a change of a specific magnitude. This minimal detectable difference relates to the research design and sample size and is the underlying concept of statistical power and sample size estimations. Therefore, it is possible to detect changes that are statistically significant but not clinically significant when the assessment techniques have very little error or when large sample sizes are used. For example, some sophisticated measures of range of motion of the hallux may have a typical error of less than 1.0°. If a sufficient number of subjects is used and there is an effect of 2°

Table 3. Relative Thresholds of Measurements Used for Clinical Interpretation of Change Score Data Sets

Measurement Threshold	Research Design/Method	Important Factors
Clinical outcome studies		Measure is in functional or quality-of-life scale
RID	Clinical changes related to the difference between the extremes of conditions	This threshold defines the range from one extreme to the other and, therefore, can identify the number of incremental MCID changes resulting in an excellent outcome
MCID	Clinical efficacy studies—RCTs Outcome measures are function or quality-of-life scales The magnitude of the MCID is determined from scores of patients who report a minimal improvement or decrease in function	The generalizability of the MCID is related to patient factors and is context specific
Reliability studies		
Minimal detectable difference	Important for biomechanical assessments and mechanistic studies using impairments Derived from studies reporting reliability estimates and population variance	This is the smallest difference the researcher should use to determine the sample size required to undertake the study Measurement is any type of scale, ie, impairment or functional
Typical error	Reliability study designs vary according to the major source of error that is being quantified/investigated	Three sources of error: • instrument • person conducting the test • populations tested

Abbreviations: RID, really important difference; MCID, minimal clinically important difference; RCT, randomized controlled trial.
Note: RID > MCID > minimal detectable difference ≥ typical error.

(on average), this value may reach statistical significance, yet such a change may be considered clinically unimportant or trivial.

To define the minimal clinically important difference, the researcher needs to interact with real patients in the real clinical setting. The perception of minimal improvements depends on many factors, including the patient's experience, expectations, and social context and the scales being used.[15] A change score of a lesser magnitude than what the patients considered the minimal important change may be considered trivial for that patient population.

If the threshold for a minimal clinically important difference for improvement and decline in status is known, then the clinician, given the mean and confidence limits of the change scores, can estimate the proportion of scores likely to lie outside the "nontrivial" threshold. This estimation of the likelihood of a good outcome to a no-change or poor outcome may be the key factor in advising the patient on a specific clinical course of action. For example, Figure 1 shows the range of changes (95% confidence intervals) of the impact of orthoses on rearfoot values, and the shaded region reflects a hypothetical range of trivial responses defined by the typical error (see Hopkins[13] for a review of typical error). From this figure, it is clear that, on average, the change due to wearing the orthosis is approximately 0.85°, with the upper confidence limit approximately 2°. A large proportion (one-third) of the sample has changes that are within the measurement error of 0.5°. From this population the clinician can be equally confident that few, if any, subjects will have a measurable decrease in rearfoot angle.

Measuring What Is Important to the Patient

In the previous example, we see an interpretation of the changes in a rearfoot angle attributable to an orthosis. Certain rearfoot angles may be considered to be the impairment and can be used to investigate biomechanical mechanisms for diagnosis or prognosis and clinical decision making; however, they are not valid in assessing treatment efficacy. The assessment of range-of-motion impairments is a common focus of podiatric medical research. From the patient perspective, their minimal clinically important difference may be 2°, and, therefore, only a small proportion of patients could be expected to detect any clinical difference. However, unless the impairment correlates strongly with functional outcomes or quality-of-life scales (eg, Landorf and Keenan[16]), the use of an impairment to document treatment efficacy is inappropriate. Similarly, demonstrating a mechanical adaptation may not necessarily translate to a statistically significant or clinically meaningful change in the patient's function. Many podiatric medicine articles investigate changes in impairments associated with biomechanical modifications of the foot posture (orthoses) and provide discussions and conclusions translating the changes in impairment to the probable or possible changes in functional outcomes. The latter is an area that warrants further investigation, including quantification of the strength of associations between various impairments and actual functional outcomes across various clinical populations.

In terms of population health, there have been recent arguments that the minimal clinically important difference represents the smallest quantum of change in the individual's perceived status. It is unclear how many times this small increment needs to be achieved to represent a complete recovery from the condition. Therefore, a larger threshold of difference has been identified as a "really important difference."[17] This is the magnitude of change on a scale between the extremes in a patient group continuum. In a population of patients with rheumatoid arthritis, the really important difference measures were found to be approximately three to six times larger than the minimal clinically important difference, and, therefore, interventions can be compared relative to the differences between highly disabled and complete recovery.

In summary, clinicians have a range of threshold changes in outcome measures that they may consider detectable, minimally clinically important, or really important according to how they choose to interpret the results. Reporting the *P* value in a clinical study does not provide the information to allow this interpretation to occur.

Conclusion

The *P* value tells the clinician about the probability of the data being produced by random changes in the scores. Highly significant *P* values do not reflect the magnitude of the change, and if a difference is not detected at a specific *P* value, then the power of the test will need to be greater if a significant *P* value is to be found in subsequent studies. This can be achieved in various ways that should have been considered at the time of research planning.

Clinicians (and researchers) may use their understanding of the typical error of an assessment protocol to determine the magnitude of change likely to be detected according to the subject numbers and expected treatment effect size. To report the treatment efficacy, however, researchers may assess the proportion of patients who exceeded the minimal clinically

important difference (or other threshold they see as valid) on scales that reflect function or participation rather than impairments.

Reports of change scores allow the podiatric physician to develop clinical hypotheses on the likelihood of the mean patient response above a specific nontrivial threshold of change (beneficial or harmful). This information is applicable to the clinician, and, therefore, researchers undertaking clinical studies should strive to provide clinical relevance with a focus on the magnitude of the change scores. The magnitude and distribution of the changes may allow some level of clinical interpretation of how these results may affect a similar patient receiving a similar intervention despite the exact *P* value and the respective level of statistical confidence.

Financial Disclosures: None reported.
Conflict of Interest: None reported.

References

1. Stern JM, Simes RJ: Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ **315:** 640, 1997.
2. Olson CM, Rennie D, Cook D, et al: Publication bias in editorial decision making. JAMA **287:** 2825, 2002.
3. Dickersin K, Olson CM, Rennie D, et al: Association between time interval to publication and statistical significance. JAMA **287:** 2829, 2002.
4. Krzyzanowska MK, Pintilie M, Tannock IF: Factors associated with failure to publish large randomized trials presented at an oncology meeting. JAMA **290:** 495, 2003.
5. Ioannidis JP: Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA **279:** 281, 1998.
6. Redmond A: Two feet or one person? The Foot **14:** 1, 2004.
7. Menz HB: Two feet, or one person? Problems associated with statistical analysis of paired data in foot and ankle medicine. The Foot **14:** 2, 2004.
8. Menz HB: Analysis of paired data in physical therapy research: time to stop double-dipping? J Orthop Sports Phys Ther **35:** 477, 2005.
9. Altman DG, Bland JM: Statistics notes: units of analysis. BMJ **314:** 1874, 1997.
10. Altman DG, Bland JM: Absence of evidence is not evidence of absence. BMJ **311:** 485, 1995.
11. Marks HM: Rigorous uncertainty: why RA Fisher is important. Int J Epidemiol **32:** 932, 2003.
12. Young KD, Lewis RJ: What is confidence? part 2. Detailed definition and determination of confidence intervals. Ann Emerg Med **30:** 311, 1997.
13. Hopkins WG: A new view of statistics. Available at: http://www.sportsci.org/resource/stats/index.html. Accessed April 9, 2005.
14. Bland JM, Altman DG: Measuring agreement in method comparison studies. Stat Methods Med Res **8:** 135, 1999.
15. Beaton DE, Boers M, Wells GA: Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol **14:** 109, 2002.
16. Landorf KB, Keenan AM: An evaluation of two foot-specific, health-related quality-of-life measuring instruments. Foot Ankle Int **23:** 538, 2002.
17. Wolfe F, Michaud K, Strand V: Expanding the definition of clinical differences: from minimally clinically important differences to really important differences: analyses in 8931 patients with rheumatoid arthritis. J Rheumatol **32:** 583, 2005.