Level V Evidence

Research Pearls: The Significance of Statistics and Perils of Pooling. Part 1: Clinical Versus Statistical Significance

Joshua D. Harris, M.D., Jefferson C. Brand, M.D., Mark P. Cote, P.T., D.P.T., M.S.C.T.R., Scott C. Faucett, M.D., M.S., and Aman Dhawan, M.D.

Abstract: Patient-reported outcomes (PROs) are increasingly being used in today's rapidly evolving health care environment. The value of care provision emphasizes the highest quality of care at the lowest cost. Quality is in the eye of the beholder, with different stakeholders prioritizing different components of the value equation. At the center of the discussion are the patients and their quantification of outcome via PROs. There are hundreds of different PRO questionnaires that may ascertain an individual's overall general health, quality of life, activity level, or determine a body part-, joint-, or disease-specific outcome. As providers and patients increasingly measure outcomes, there exists greater potential to identify significant differences across time points due to an intervention. In other words, if you compare groups enough, you are bound to eventually detect a significant difference. However, the characterization of significance is not purely dichotomous, as a statistically significant outcome may not be clinically relevant. Statistical significance is the direct result of a mathematical equation, irrelevant to the patient experience. In clinical research, despite detecting statistically significant pre- and post-treatment differences, patients may or may not be able to perceive those differences. Thresholds exist to delineate whether those differences are clinically important or relevant to patients. PROs are unique, with distinct parameters of clinical importance for each outcome score. This review highlights the most common PROs in clinical research and discusses the salient pearls and pitfalls. In particular, it stresses the difference between statistical and clinical relevance and the concepts of minimal clinically important difference and patient acceptable symptom state. Researchers and clinicians should consider clinical importance in addition to statistical significance when interpreting and reporting investigation results.

From the Houston Methodist Orthopedics & Sports Medicine, Institute of Academic Medicine, Houston Methodist Research Institute (J.D.H.), Houston, Texas; Weill Cornell Medical College (J.D.H.), New York, New York; Heartland Orthopedic Specialists (J.C.B.), Alexandria, Minnesota; UConn Musculoskeletal Institute, Human Soft Tissue Research Laboratory UConn Health (M.P.C.), Farmington, Connecticut; The Centers for Advanced Orthopaedics, The Orthopaedic Center (S.C.F.), Rockville, Washington DC; and Penn State Hershey Bone and Joint Institute (A.D.), Hershey, Pennsylvania, U.S.A.

The authors report the following potential conflicts of interest or sources of funding: J.D.H. receives support from Smith & Nephew, NIA Magellan, Depuy Synthes, and SLACK; and is a board member of Arthroscopy, *AANA Research Committee, AOSSM Self-Assessment Committee, and AAOS OA Performance and Function Workgroup. J.C.B. is a board member of* Arthroscopy. *S.C.F. receives support from Smith & Nephew, Ceterix, Synthes, and Ossur; and is a board member of* Arthroscopy, *AOSSM, and ISAKOS. A.D. receives support from Biomet and Smith & Nephew; and is a board member of* Arthroscopy, *AANA, and OJSM.*

Received October 28, 2016; accepted January 23, 2017.

Address correspondence to Joshua D. Harris, M.D., Houston Methodist Orthopedics & Sports Medicine, 6550 Fannin Street, Smith Tower, Suite 2500, Houston, TX 77030, U.S.A. E-mail: cristyhayes@comcast.ne

0749-8063/161051/$36.00

http://dx.doi.org/10.1016/j.arthro.2017.01.053

Clinical studies are increasingly using patient-reported outcome (PRO) measures to quantitatively capture the effect of an intervention. Statistical analysis quantifies the size and precision of the differences in PRO scores before and after an intervention or between groups of patients. Statistical methods are blind to the clinical relevance. In other words, "patients don't know what their *P* value is, nor do they care." Consequently, statistical significance may or may not reflect a clinically meaningful change. A patient should be able to perceive the effect of the intervention (i.e., treatment) as "better" or "improved" if it is clinically meaningful. Furthermore, this difference should meet a minimum threshold of satisfaction ("I'm happy," "I'm satisfied," or "I'd do the intervention all over again").

A recent Level 1 evidence randomized controlled trial comparing autologous chondrocyte implantation (ACI) and microfracture in the knee at 3-year follow-up has been published by Saris et al.,[1] which highlights the difference between statistical significance and clinical relevance. The authors showed statistically significant

($P < .05$) differences favoring ACI in the overall Knee Injury and Osteoarthritis Outcome Score (KOOS) and 2 KOOS subscores (score range, 0-100). The "overall" KOOS difference between the 2 groups was only 2.3 points (77.6 vs 75.3). However, can a patient and his/her physician detect a difference between a KOOS of 77.6 and 75.3? After anterior cruciate ligament reconstruction, a change in a KOOS of 8 to 10 has been suggested to denote the minimal clinically important difference (MCID) detectable by a patient.[2] This indicates that a patient is unlikely to perceive a difference of 2.3 points, despite the statistically significant difference suggesting such. Thus, it is the responsibility of authors, journal editors, and readers themselves to ensure that the interpretation and clinical translation of an investigation's conclusions meet not only statistical but also clinically meaningful thresholds. This requires all participants in peer review to critically analyze study results and conclusions.

Statistical Basis for Clinical Relevance

Distinguishing between clinical and statistical significance requires an understanding of the role of a random error. A random error is variability around the true mean in the outcome being measured. The amount of posterior tibial slope, dimensions of the supraspinatus insertion, and preoperative pain levels are examples of objective measures that differ between individuals due to inherent biological variation. It is impossible to know the true mean of the population as that would require measuring every single individual. Thus, samples of the population are used and summarized statistically with means and measures of variation (standard deviation) around the mean. Sampling, however, could include individuals who are away from the mean, "skewing" the data toward an erroneous mean. Increasing sample size captures a larger portion of the population, improving precision in estimating the true mean and reducing the effect of the random error. In this regard, all study results are impacted by the random error to some degree. This is an important point as the random error has historically been treated as either present or absent.

Specifically, null hypothesis testing and the corresponding probability or "*P*" values dichotomize the effect of the random error to "significant" or "not significant." This fundamentally flawed approach can lead to the conclusion that a particular finding that does not cross the threshold of 0.05 is neither real nor important. The size of the difference and how precisely it has been estimated should be of interest rather than a yes or no decision as to whether a difference exists. Confidence intervals are a useful measure for determining how precisely a difference has been estimated and are a preferred measurement of the random error by *Arthroscopy*. The upper and lower limits of the interval and how wide or narrow they are provide a measure of precision. Narrow limits reflect greater precision in estimating the difference, thereby reducing the random error.

Type I Error and Type II Error

Interpretation of an investigation's results requires more judgment than simply determining the presence or absence of a statistically significant difference. The terms "type I error" and "type II error" are often used to discuss the risk of misinterpreting the results of a study. These errors are the result of testing the null hypothesis and need to be considered. For example, when comparing 2 or more samples, researchers first designate a null hypothesis. The null hypothesis generally states that the samples or groups being studied are not different from each other, in a superiority design study. If the results of the study have means that are different enough from each other in their size and variance, then comparing these samples will result in "statistical significance."

A type I error occurs when the null hypothesis is discarded despite it being true. In other words, the difference observed between groups is assumed and reported to be true when, in fact, the difference does not actually exist, particularly a problem in investigations with a large "n," such as "big data" research. When the results of a study are statistically significant, a type I error should be considered. Conversely, researchers can also make an error by stating that no difference exists between groups when there truly is a difference. This is a type II error. A type II error is much more common than type I and is a risk with small sample size studies that failed to detect significant differences between groups. In this situation, the alternative hypothesis is true rather than the null hypothesis.

The investigation's power is crucial to determine the impact of both significant and nonsignificant results. The ability of a study to detect a difference when one truly exists is referred to as statistical power, defined as 1 minus the type II error rate. Traditionally, investigators aim to enroll enough subjects to reach at least 80% power. This means that if a difference truly exists, the study will detect it 80% of the time (20% type II error rate). Some investigations may be underpowered to detect a difference between groups if one existed. Online power calculators are readily and freely available. The calculation of an adequate sample size requires the type I error rate (typically 0.05), desired power (typically 80%), the size of the difference the investigators are trying to detect (quantitatively, should be at least the MCID—optimizes collinearity of statistical significance and clinical importance), and an estimate of variation (standard deviation) to calculate the required sample size. Some calculators require effect