Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

CrossMark

Editorial

Editorial: The Minimum Clinically Important Difference—The Least We Can Do

Seth S. Leopold MD, Raphaël Porcher PhD

H ow should a clinician decide whether a treatment effect claimed in a journal article is worth caring about?

Does the answer lie in a p value? Or instead, should the reader focus on whether the size of the treatment effect exceeds the minimum clinically important difference (MCID), in the hopes of determining whether the treatment's benefit is worth its risks and costs?

Though all of these statistical analyses matter, the conversation too often starts and stops with the p value. Such a superficial approach results in patients enduring treatments that may not deliver benefits large enough for them to perceive, and far too small to justify the risks or costs involved. A more-nuanced approach is called for.

Some have called for journals to abandon the p value and its related concepts completely [3, 15]. This rather-extreme viewpoint is unlikely to catch on since readers need some means to evaluate how likely it is that chance could have played a role in obtaining the observed results. We do agree, though, that clinicians should interpret p values with greater care than usually is exercised [16]. Clinicians should be open to approaches apart from frequentist statistics (p values and the like) for this purpose. Alternatives range from simple common sense [7] to more-sophisticated Bayesian statistics [16], which can help the reader arrive at a more-complete understanding of the data.

But since most claims in orthopaedic research papers still end with p values, $CORR^®$ will ask authors to use those p values thoughtfully. We suggest that authors set sensible a priori thresholds for p values based on the experiment itself, before analyzing any data. No law mandates a threshold of 0.05. In fact, it often seems to us that higher or lower thresholds would be reasonable (the former for exploratory studies or studies involving little risk of harm, and the latter for studies that propose interventions that carry larger risk or greater potential for toxicity, and studies in which many statistical analyses are performed), though most studies embrace the near-mystical 0.05 with both arms. And once an investigator has decided on reasonable thresholds, it is best not to play fast and loose with them after seeing the data. Claims of "nonsignificant differences" or "trends" about findings the author had hoped to see but did not injects an additional element of subjectivity that is best omitted; one seldom sees authors reporting trends about findings that disagree with investigators' preconceived notions.

A small p value is far from the whole story. Large studies can detect small differences between treatment approaches with a high degree of "statistical significance" (that is, small p values). All too often, authors use

S. S. Leopold MD (✉)
Clinical Orthopaedics and Related Research, 1600 Spruce St., Philadelphia, PA 19103, USA
e-mail: sleopold@clinorthop.org

R. Porcher PhD
Center of Clinical Epidemiology, Hôpital Hôtel-Dieu; Team METHODS, Centre de Recherche en Epidémiologie et Statistiques (CRESS), INSERM U1153, Paris Descartes University, Paris, France

Editorial

those small p values—numbers like 0.001 or less—to suggest "there is a huge difference" when all that finding might mean is "we're fairly sure there is a difference." Sometimes those differences are imperceptible (or hardly perceptible) to our patients. Let's agree that in those instances, the intervention is not worth enduring.

This is where the concept of the MCID [10] and its close relative, the minimum detectable change (MDC), come into play: How large must a treatment effect be for a patient to detect it, or care about it? The answer will depend on the condition being treated, the patient population undergoing the treatment, and the outcomes tools used to measure the results [4, 11]. And even where all those factors are held constant, there are (generally) two major ways to calculate an MCID: Anchor-based approaches, and distribution-based approaches. The latter are easier to come by, since all one needs is a dataset and a calculator; for example, MCIDs can be estimated as a function of the standard deviation of the data, and some studies find this approach to be robust [13]. But many observers [5, 14]—including the authors [12] of the review in this month's Statistics in Brief article in *Clinical Orthopaedics and Related Research*®—believe that anchor-based calculations of the MCID, which define "clinically important" in

relation to changes identified as important by the patients themselves, are more relevant to clinical practice. Although this approach can seem subjective, ultimately it is the patient's perception that matters most, and anchoring a MCID to a difference that patients have defined as important makes the most sense to us, as well.

Authors analyzing surgical treatments should present results in terms of MCID whenever it is practical to do so. In this month's *CORR*®, Maltenfort and Diaz-Ledezma [12] provide the MCIDs for dozens of the most-commonly used orthopaedic outcomes tools. Studies using outcomes tools for which no MCIDs are available should consider suggesting how large the effects would need to be for them to matter to patients, and they should justify those contentions.

Evaluating the treatments we use in terms of MCIDs is important; differences smaller than the MCID are unlikely to matter much to patients, and certainly are not worth paying for with dollars or surgical risk. A miniscule p value ($p < 0.001$) attached to an effect size smaller than the MCID, is, by definition, a "difference" that a patient is unlikely to call important. The larger the sample size, the more likely we are to identify such "statistically significant" but clinically unimportant differences. As more studies draw from registries, insurance

databases, or national quality-improvement repositories like the Nationwide Inpatient Sample, or the National Surgical Quality Improvement Project—all of which summarize the experiences of vast numbers of patients—the issue of proving clinical relevance in addition to statistical significance becomes even more important.

We note that sometimes an even more subtle approach is necessary. When the mean effect size for some intervention is below the MCID, it remains possible that a subset of patients may have benefitted from treatment. There is nothing wrong (and there may be a lot right) with analyzing individual-patient data when possible to determine whether a subset of patients experienced a clinically important benefit from that intervention, and if they had anything in common. Doing so might identify subgroups of patients that could benefit from future research on that intervention, even if the overall population, in aggregate, did not benefit. But the big picture remains this: Patients see the world in terms of effect sizes, not p values; it is time for clinicians, research scientists, and medical editors to do likewise.

The MCID is the least we can do; in fact, before we would consent to any procedure involving serious risks or costs, we would ask for more.

Editorial

Specifically, we would look for some evidence that the procedure's impact on our health or quality of life would be more than "minimal." Presumably, our patients feel likewise. Some authors have suggested that for large interventions—like surgery—patients deserve something more substantial than the MCID. Alternatives include substantial clinical improvement [9], minimum acceptable outcomes [2], improvement as a percentage of possible improvement [8], and others [6], but as yet, these have not taken hold. Until or unless they do, the MCID is what we have, and it is only after establishing that a procedure, drug, or device delivers at least a clinically important improvement that we can begin to ask second-order questions, like whether that benefit is worth the costs or risks involved with its use. Certainly those that do not, are not.

A few years ago, an editor whom we respect wrote that MCIDs will soon "become a historical oddity, like articles that blissfully judged an operation's success based entirely on a surgeon's impression of his own good work" [1]. He envisioned a post-MCID world where still-more-refined mechanisms will be devised to ascertain whether an intervention justifies its risks and costs. Sadly, most orthopaedic research is presented as though we practice in a pre-MCID world. A disconcerting number of the papers we evaluated last year drew inferences from "differences" too small to be statistically detectable, and many others described sub-MCID differences as "significant," hoping that readers would fail to distinguish between statistical significance and clinical relevance.

We can tell the difference, and it is important we try to do so.

Where possible, we will continue to ask authors to present their differences in terms of MCIDs or other metrics that help describe the clinical relevance of their findings, and not to make grandiose claims about the benefits of treatments that are merely statistically detectable (or less). To help readers and writers, we recommend the overview on the topic in this issue, which includes handy reference tables of MCIDs for each specialty within orthopaedics for which they are available [12].

References

1. Carragee EJ. The rise and fall of the "minimum clinically important difference." *Spine J.* 2010;10:283–284.
2. Carragee EJ, Cheng I. Minimum acceptable outcomes after lumbar spinal fusion. *Spine J.* 2010;10:313–320.
3. Colquhoun D. The problem with p-values. *Aeon.* 11 October 2016. Available at https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant, accessed on December 29, 2016.
4. Copay AG, Martin MM, Subach BR, Carreon LY, Glassman SD, Schuler TC, Berven S. Assessment of spine surgery outcomes: Inconsistency of change amongst outcome measurements. *Spine J.* 2010;10:291–296.
5. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol.* 2003;56:395–407.
6. Davis AM, Perruccio AV, Lohmander LS. Minimally clinically important improvement: All non-responders are not really non-responders: An illustration from total knee replacement. *Osteoarth Cartilage.* 2012;20:364–367.
7. Dorey F. Statistics in brief. Interpretation and use of p values. All p values are not equal. *Clin Orthop Relat Res.* 2011;469:3259–3261.
8. Gilmer BB, Comstock BA, Jette JL, Warme WJ, Jackins SE, Matsen FA III. The prognosis for improvement in comfort and function after the ream-and-run arthroplasty for glenohumeral arthritis. An analysis of 176 consecutive cases. *J Bone Joint Surg Am.* 2012;94:e102.
9. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. *J Bone Joint Surg Am.* 2008;90:1839–1847.

Editorial

10. Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. *Cont Clin Trials.* 1989;10:407–415.

11. Katz NP, Paillard FC, Ekman E. Determining the clinical importance of treatment benefits for interventions for painful orthopaedic conditions. *J Ortho Surg Res.* 2015;10:24.

12. Maltenfort M, Diaz-Ledezma C. Statistics in brief: Minimum clinically important difference—Availability of reliable estimates. *Clin Orthop Relat Res.* [Published online ahead of print January 3, 2017]. DOI: 10.1007/s11999-016-5204-6.

13. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care.* 2003;41:582–592.

14. Rai SK, Yazdany J, Fortin PR, Aviña-Zubieta JA. Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis Res Ther.* 2015;17:143.

15. The B. Significance testing—Are we ready yet to abandon its use? *Current Med Res Opin.* 2011;27:2087–2090.

16. Wasserstein RL, Lazar NA. The ASA's Statement on p-values: Context, process, and purpose. *Am Stat.* 2016;70:129–133.