All interventions differ, although some are more different than others

Parsons NR[(1)], Perry DC[(2)] and Costa ML[(3)]

Correspondence to N. R. Parsons; email: nick.parsons@warwick.ac.uk

It is a well-known statement of statistical wisdom that all treatments differ.[1] We could substitute the word 'intervention' or the phrase 'surgical procedure' for 'treatment' and qualify the word 'differ' by, for instance, saying "have different outcomes". Yet, in doing so we challenge far more than semantics, questioning the relevance of much of the research reported in *The Bone & Joint Journal*, and the wider medical literature. For example, the DRAFFT study[2-4] compared the clinical effectiveness of Kirschner wire fixation with locking plate fixation for patients with a dorsally displaced fracture of the distal radius. Selectively picking out the key result of the paper one would conclude that there was no difference in outcomes between the two groups; thus that there was no difference between the treatments. This result seems to contradict the first statement in this paragraph, which follows from the simple argument that by definition all treatments must be different because no two things can ever be exactly the same in every respect. This sounds like a matter of philosophical interest only, but understanding how we reconcile this apparent contradiction helps shed light on an issue that continues to be a cause of confusion for many authors submitting papers to our Journal.

The arguments we use when formulating, testing, reporting and, particularly, interpreting the results of statistical analyses, are widely misunderstood but crucial to understanding much published clinical research. When reporting the results of studies which compare outcomes between treatment groups or experimental units, we almost always test for a significant difference between groups using a statistical test that conveniently provides a p-value to aid interpretation. The ubiquitous p-value cut-off of 0.05 allows us to classify all results as either significant or non-significant. This objective criterion, which we call statistical significance, has proved to be simple and effective, despite much criticism.[5] However, a fundamental aspect of this methodology is that the smallest imaginable difference between treatment groups can be shown to be statistically significantly different simply by increasing the sample size, typically the number of subjects in the study. As we increase the sample size we get progressively more precise estimates of measures such as the mean, so eventually we find that, however small the difference is between means, it will be deemed to be statistically significantly different. So we can always show, if we are prepared to collect enough data, that any treatments we test must be different as we can always, in principle, provide evidence of statistical significance; i.e. all treatments differ. Clearly such small differences are likely to be of no clinical importance; therefore, evidence to support a statistically significant difference does not necessarily imply that the difference is clinically important.

In order to both plan and interpret the results of all clinical research, we must use the concept of a minimum clinically important difference (MCID); sometimes referred to as minimum important difference (MID). This is the magnitude of difference, or change, in outcome that is considered to be important, either beneficial or harmful, or relevant to the patient, that would lead the patient or clinician to consider a change in treatment. By relating all reported differences in treatment to the MCID, we are able to differentiate between clinical significance and statistical significance. The latter does not always imply the former, but the former requires evidence of the latter for us to believe in its veracity.

Returning to the report of the DRAFFT study, a careful reading shows that the authors are clear that "…as the confidence intervals exclude the minimum clinically important

difference…we conclude that any difference…between treatment groups is unlikely to be important to patients…".[2] Because the estimated effect size was smaller than the MCID for the primary outcome measure, and the confidence interval also excluded the MCID, the conclusion was that the difference between treatment groups was not clinically important. If the study sample size had been doubled or trebled, we may well have found evidence of statistical significance, but, all other things being equal, we would still have concluded that the differences were too small to be clinically important. This explains why the MCID for a selected outcome measure is crucial in the design of much clinical research and, in particular, randomised controlled trials (RCTs). When undertaking a power analysis before starting an RCT, a key determinant of the sample size is the MCID. Research studies are typically powered to detect effects that are as small as the MCID, but no smaller, with high probability. This helps to moderate or limit the sample size so that we do not waste resources collecting such a quantity of data that we can estimate tiny effects, which are of no clinical importance, with high precision.

How do we determine the MCID for an outcome measure? For familiar (hard) outcomes such as weight or blood pressure, we can all naturally at least approximate the likely range of the MCID; e.g. an improved weight loss for a new dieting intervention, against a standard, of 5 kg would clearly be clinically important, whereas a change of 10 g would be unimportant. However, the definition of a MCID is generally more problematic for patient-reported outcome measures (PROMs) such as function, quality of life or pain. Briefly, there are two main approaches to determine MCIDs for PROMs:[6] anchor-based, or distribution- based methods. Anchor-based methods compare the change in a PROM score to another external measure of change, the anchor. Most commonly this anchor is another subjective assessment by which the patient, for instance, rates their condition as much better, better, unchanged, worse, or much worse. Distribution-based methods compare the change in a PROM to a measure of variability such as an effect size or standard deviation (SD). The one-half SD method considers that a patient improving by more than one-half of the SD of an outcome measure has achieved a MCID. For many important and widely used PROMs such as the Oxford Hip Score (OHS) and Oxford Knee Score (OKS), MCIDs are well established and widely accepted.[7]

Authors submitting manuscripts to The Bone & Joint Journal must be precise in how they report significance and distinguish between clinical and statistical significance when drawing conclusions. It is not always sufficient simply to answer the question "are the treatments different?" based on p-values and evidence of statistical significance. If the answer to this question is "yes", then one must always follow-up by asking "how different?"; before proceeding to report confidence intervals and to compare estimated effect sizes to the MCID – only then can we truly understand and assess the importance of the research.

References:

1. Nester MR. An applied statistician's creed. Appl Stat 1996;45:401–410.
2. Costa ML, Achten J, Parsons NR, et al. Percutaneous fixation with Kirschner wires versus volar locking plate fixation in adults with dorsally displaced fracture of distal radius: randomised controlled trial. BMJ 2014;349:4807.
3. Tubeuf S, Yu G, Achten J, et al. Cost effectiveness of treatment with percutaneous Kirschner wires versus volar locking plate for adult patients with a dorsally displaced fracture of the distal radius: analysis from the DRAFFT trial. Bone Joint J 2015;97-B:1082–1089.
4. Costa ML, Jameson SS, Reed MR. Do large pragmatic randomised trials change clinical practice?: assessing the impact of the Distal Radius Acute Fracture Fixation Trial (DRAFFT). Bone Joint J 2016;98-B:410–413.
5. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat 2016;70:129–133.

6. Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7:541–546.
7. Beard DJ, Harris K, Dawson J, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol 2015;68:73–79.

Affiliations:
(1) University of Warwick,
Statistics and Epidemiology,
Warwick Medical School,
Coventry CV4 7AL, UK.
(2) University of Liverpool,
Institute of Translational
Medicine, Easton Rd, Liverpool
L12 2AP, UK.
(3) University of Oxford,
NDORMS, Oxford Trauma,
Kadoorie Centre, John Radcliffe
Hospital, Oxford, OX3 7UD, UK.