

Correct reporting and interpretation of clinical data

“As the p value was smaller than 0.05, we conclude that the treatment was effective”.

Such conclusions are often made, and the treatment concerned is considered effective from a statistical point of view. However, the magnitude of the treatment’s effectiveness may not be meaningful or important: although “real”, the size of the difference is too small to matter to patients or clinicians. Covering this and other issues, the American Statistical Association has published a policy statement on p -values (Wasserstein and Lazar, 2016): a small p value often induces the inappropriate conclusion of a large treatment effect. The p value does not measure the magnitude of a treatment effect or the importance of the results. It is just the probability under a specified statistical model. The calculation of p values is influenced by the sample size. The larger the sample, the greater the likelihood of a “significant” p value, even if there is only a small effect of a given treatment. This may also occur during meta-analysis, in which the results of multiple studies are pooled statistically, increasing the sample size.

In contrast, two undesirable scenarios may occur with a small sample size in a study:

1. Statistical “significance” might not be demonstrable with too small a sample, despite a large treatment effect existing, because of this relationship between the sample size and the power of the statistical analysis. This scenario is known as a type 2 error.
2. If an apparently large difference is reported based on a small sample size without any supporting statistical analysis, it may be assumed that the large difference must be a true effect, and has not arisen by chance. This is not necessarily the case, yet this latter scenario is also commonly encountered.

Therefore, it is important that appropriate statistical analysis is undertaken, and that a calculation is performed during the design of a study to determine the optimal sample size. In situations where sample size calculation is not possible before execution of the study, a power calculation might be performed after the investigators have captured the data, to describe the risk of a type 2 error in the statistical analysis. Such calculations traditionally use effect sizes. However, using clinically-meaningful values would make the analysis more applicable to what matters in the real world. Guidance on the specification and reporting of target differences in trials has been provided by the Difference Elicitation in Trials (DELTA) project (Cook et al., 2015).

In recent years, interpreting clinical study results based on what is important to the patients has become increasingly popular. This concept is akin to “biological significance”. As an aspect of the behaviour of an outcome measure, this is referred to as “interpretability”. Suitable interpretability parameters include the minimal important change (MIC) and the minimal important difference (MID). These terms have important differences. The MIC is the smallest change that a patient perceives as important, whereas the MID is the smallest difference between patients considered important (Mokkink et al., 2010).

MICs aim to define important change for an individual, as would be observed following treatment in a cohort study. They can be calculated using either distribution-based or anchor-based methods. Distributional methods compare the mean change to a measure of its variability, such as the standard deviation or standard error of measurement. Anchor-based methods categorise the change against an external factor, which is used to define improvement. One potential advantage of anchor-based methods is that they can account for what matters to patients (or clinicians), depending on the anchor chosen. Transition items are often used as anchors. For example, at follow-up, the patient may be asked to

reflect upon their treatment using a scale of from success to failure (de Vet et al., 2007). The response to the transition item is used to categorise patients into those who have experienced benefit and those who have not, and then an MIC estimate can be calculated using the mean change within the categories, by a receiver operating characteristic (ROC) curve, or by regression modelling.

MIDs define the importance of the difference between groups, and so can be used to support the comparison of different treatments, in randomised controlled trials, for example (de Vet et al., 2006; Mokkink et al., 2010). The MID is typically determined by the difference between the mean change score of patients who experienced anchor-defined benefit and the mean change score of those who did not. MIDs were previously called minimal clinically important differences (MCIDs).

The “subjectivity” of anchor-based MICs and MIDs may be considered a weakness, yet it is also strength. Patients’ varied interpretations of “successful” treatment when responding to an anchor question may seem unreliable. However, patients’ interpretations of outcomes might be expected to correlate with their behaviour beyond the health system (e.g. whether they will return to work, or whether they need social care at home). In such a situation, it can be argued that patients’ “subjective” opinions are of greater relevance than clinicians’ potentially arbitrary definitions of treatment success and failure, which might not relate to real-world improvement.

Some examples of interpretability estimates from recent studies are listed in Table 1, and earlier interpretability estimates in elective surgery of the hand are consolidated in a recent review article (Rodrigues et al., 2015). It is notable that interpretability estimates are context-specific: they may vary between outcome measures, conditions, treatments and populations. Logical explanations for this exist. For example, patients may need to experience greater improvement following high-risk or invasive treatments to consider the

treatment worthwhile. Furthermore, different methods exist to calculate interpretability parameters like MICs, hence they are reported as 'estimates' or ranges (Revicki et al., 2008). Where available, consensus on values to use as an MIC or MID (achieved by processes such as Delphi studies) may be useful. Unfortunately, this is not yet available for many hand conditions.

In hand surgery research, MICs and MIDs have been used infrequently in comparison to other disciplines. Table 2 includes papers published in this journal in which the authors considered interpretability estimates for either the sample size calculation and/or interpretation of their study results. In keeping with the opening quote of this article, the interpretation of the results often differs depending on whether it is based on statistical significance or clinical meaningfulness. Examples are discussed in Table 2.

In summary, similar errors often occur in reported data. These include attributing clinical meaningfulness to a statistically significant result, failing to identify a real difference due to under-powering, or failing to support conclusion with appropriate statistical analysis. Instead, the appropriate statistical analysis should be specified and used, along with data supporting the sample size or power of the study, and these data should demonstrate how the study would be able to demonstrate a clinically meaningful difference or change, if one exists. Readers should critically appraise the choice of statistical analysis in articles, and interpret the results in terms of both statistical significance and clinical meaningfulness. The evidence underpinning interpretability estimates used (such as MICs and MIDs) should be considered. Where sample size and power calculations are performed by researchers or read by the audience, the basis for the expected difference between treatments should be scrutinised.

Miriam Marks

100 Department of Teaching, Research and Development

101 Schulthess Klinik, Zurich, Switzerland

102 **Jeremy N Rodrigues***

103 Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences

104 (NDORMS), University of Oxford, Oxford, UK

105 *Corresponding author: j.n.rodrigues@doctors.org.uk

106

References

- Cook JA, Hislop J, Altman DG et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials*. 2015, 16: 12.
- de Vet HC, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. *J Rheumatol*. 2006, 33: 434-35.
- Mokkink LB, Terwee CB, Knol DL et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010, 10: 22.
- Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008, 61: 102–9.
- Rodrigues JN, Mabvuure NT, Nikkhah D, Shariff Z, Davis TR. Minimal important changes and differences in elective hand surgery. *J Hand Surg Eur*. 2015, 40: 900-12.
- Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. 2016, 70: 129-33.

Table 1: Recent minimal important changes (MICs) and minimal important differences (MIDs) for patient-reported outcome measures (PROMs) for hand disorders

Study	Population	Intervention	PROM	MIC/MID (range)
Rodrigues et al. J Hand Surg Eur. 2017, 42: 301-9.	Dupuytren's disease	Surgery	URAM	MIC: 10.5 MID: 8
Clement et al. J Hand Surg Eur. 2016, 41: 624-31.	Carpal tunnel syndrome	Carpal tunnel decompression	quickDASH	20
Marti et al. J Hand Surg Eur. 2016, 41: 957-62.	Carpal tunnel syndrome	Carpal tunnel release	EQ-5D-5L	0.09
Wehrli et al. J Hand Surg Am. 2016, 41: 896-902.	Dupuytren's disease	Collagenase injection or surgery	Brief MHQ	7
Maia et al. SICOT J. 2016, 2: 32.	Various hand / wrist conditions	Surgery	DASH MHQ	15 – 19 15
Smith-Forbes et al. J Hand Ther. 2016, 29: 81-8	Distal radius fracture Lateral epicondylitis Carpal tunnel syndrome	Surgery Non-surgical Carpal tunnel release	quickDASH	25.8 15.8 18.7
Alotaibi et al. Disabil Rehabil. 2016, 38: 2469-78.	Various upper limb conditions	Occupational and physical therapy	DASH	15
Marks et al. Arthritis Care Res (Hoboken). 2014, 66: 245-52.	First carpometacarpal osteoarthritis	Conservative or surgical treatment	MHQ	17
London et al. Plast Reconstr Surg. 2014, 133: 616-25.	Atraumatic hand / forearm diagnosis	Various treatments	MHQ	10.8 (8-13)

URAM: Unité Rhumatologique des Affections de la Main Scale; DASH: Disabilities of the Arm, Shoulder and Hand Questionnaire; EQ-5D-5L= EuroQol EQ-5D-5L; MHQ: Michigan Hand Outcomes Questionnaire

Table 2: Studies considering the minimal important change (MIC) or the minimal important difference (MID) for the interpretation of their study results

Study	Population and Intervention	Data interpretation
Yoon et al. J Hand Surg Eur. 2017, 42: 260-5.	Distal radial fractures, Open reduction and palmar locking plate fixation	Based on the p value Although the authors mention the MID in their power calculation, but do not consider it in data interpretation. The results show a smaller mean difference between the groups than the MID, but p-value is statistically significantly. → Statistical and patient-centred interpretation differ
Hameso and Bland. J Hand Surg Eur. 2017, 42: 275-80.	Carpal tunnel syndrome, Steroid injection or decompression surgery	Based on the p value and the MID There is a significant difference between the groups, but it is smaller than the MID. This fact is discussed adequately. → Statistical and patient-centred interpretation differ
Marks et al. J Hand Surg Eur. 2015, 40: 927-34.	First carpometacarpal osteoarthritis, Steroid injection or surgical treatment	Based on MIC and “significance” Treatment effect was a secondary outcome. The conclusion was based on the MIC and significance without reporting a p value. → Statistical and patient-centred interpretation are equal
Cousins et al. J Hand Surg Eur. 2015, 40: 961-5.	Patients undergoing carpal tunnel decompression	Based on the p value For sample size calculation, the MIC was used. The results show no significant difference between the groups, and the difference was smaller than the MIC. The authors do not consider interpretability data in their interpretation. → Statistical and patient-centred interpretation are equal