Annotation

3 steps to improve reporting and interpretation of patient-reported outcome scores in orthopedic studies

Using the KOOS as an example

Ewa M Roos

Center for Muscle and Joint Health, University of Southern Denmark, Odense, Denmark
Correspondence: eroos@health.sdu.dk

1. Calculate the score as recommended

Patient-reported outcome measures (PROMs) are becoming more common in orthopedic studies. PROMs are questionnaires from which 1 or several scores can be calculated. It is important to calculate the score(s) according to the instructions by the developers, so as not to threaten the validity of the measure.

For the Knee injury and Osteoarthritis Outcome Score (KOOS) the 5 subscale scores for Pain, Other Symptoms, Activities of Daily Living (ADL), Sport and Recreation function (Sport/Rec), and knee-related Quality of Life (QOL) should be calculated and reported separately on a 0–100, worst to best, scale (Roos 2012). Although calculating a total KOOS score is not recommended (Roos 2012) a total KOOS score has been reported in many orthopedic papers. The KOOS subscales hold from 4 to 17 items each and summing all items across the subscales will give very different weights to the five subscales. As an example, 40% of the contribution to a total score would originate from items related to difficulty with activities of daily living. Reporting a total score threatens the validity of the KOOS for many patient groups, including younger individuals with knee injury where function during sport and recreation and knee-related quality of life are subscales of greater relevance, and older subjects with knee osteoarthritis where pain is a subscale of equal or greater relevance compared with difficulty with function during daily activities (Collins et al. 2016).

2. Report confidence intervals instead of p-values

The recommended way to report PROM scores in clinical trials is to give the mean change score with its 95% confidence interval (95% CI), outlining the uncertainty of the estimated mean change score. Measures of uncertainty (and effect sizes) are instrumental in understanding the results and in clinical decision-making. The use of p-values is discouraged by many journals, including Acta Orthopedica (Ranstam 2005, 2012). Reasons for this include p-values being blamed for "distorting readers' perception of observed results," p-values not providing a direct estimate of how likely a result is to be true or not true, the p-value not being a measure of effect size, and "readers mistaking statistical significance for clinical significance" (Bhandari et al. 2005).

3. Clinical interpretation and the appealing but devious concept Minimal Important Change (MIC)

A statistically significant difference in PROM score is not necessarily equivalent to a clinically relevant finding. To help establish clinical relevance, the concept "minimal important change" (MIC) has been introduced. The MIC value is used to determine whether the mean change found within a group over time is considered clinically relevant. Commonly it is also applied in randomized controlled trials (RCTs) to evaluate whether the mean difference in longitudinal change found between 2 groups is clinically relevant.

Interpretation of clinical relevance, and the MIC concept as such, may not be as straightforward and simple as we would wish for. In fact, studies in orthopedics, and across medical disciplines, have found that the MIC value for a specific PROM varies with definition and calculation method used (at least 14 definitions have been put forward since 1987), wording of the anchor question, and the response options used when determining the MIC with anchor-based methods (examples include if we are asking for "change" or "important change" and what cut-off is chosen, for example "a little better," "somewhat better," or "better"), patient characteristics, type of intervention undergone (if any), and time to follow-up. In summary, it is increasingly being recognized that there is no such thing as a single MIC value that is applicable for a PROM across contexts (King 2011).

For the purpose of this Annotation let us say a MIC value of 12 has been suggested for the KOOS subscale Sport/Rec in young adults following reconstruction of the anterior cruciate ligament of the knee (ACLR). The recommended way to apply this MIC is to calculate and report the proportion of individuals with a score improvement of at least 12 following ACLR. Individuals with a score improvement of at least 12 are categorized as "responders." In a comparative trial, the proportion of responders in the 2 treatment groups should be reported. If a responder analysis was the pre-specified analysis method for the trial, the proportions of responders are compared statistically. A more important aspect to consider, however, is whether the difference in proportions of responders in the 2 groups indicates a change in clinical care. For example, will a difference in responders of 10% prompt clinicians to prefer and recommend the new treatment instead of the old or is the extra 10% of responders considered to be a marginal improvement not sufficient to warrant a new treatment being introduced? And what if the difference in proportions of responders were 20% or 30%: would any of those differences in success rates make patients, clinicians, and health care administrators prefer the new treatment? This is a discussion authors need to have prior to starting a study comparing 2 treatments.

Despite some authors questioning the MIC concept as such, there is agreement that clinical interpretation of study results is of utmost importance. An alternative approach is to perform a responder analysis and report 20% and 50% improvements in the PROM score as moderate and large response rates, respectively (Felson et al. 1993, Escobar et al. 2012). This approach may also be applied in the increasing number of RCTs in orthopedics comparing interventions of different risk and cost profiles. In these studies it is not granted that similar improvements are required in the two groups to define a successful outcome. In patients with knee problems, it may be that clinicians and administrators would consider a 20% improvement in PROM score satisfactory following a low-risk and cheap intervention such as exercise therapy while a 50% improvement would be required for a satisfactory result following surgery because of its greater invasiveness, risk, and higher cost. On the other hand, it may be that patients perceive exercise as time-consuming, boring, and uncomfortable, while surgery,

hospital stay, and rehabilitation is more tolerable since it is often believed to be necessary to improve the condition. To facilitate interpretation of the results, authors (and patients) would have to decide a priori what level of improvement constitutes a success for the treatments compared and what difference in success rates between the respective treatments would indicate a change in clinical recommendation.

Commonly, including by me, the MIC has been used to evaluate whether between-group differences are clinically relevant. It is, however, no longer obvious to me and others (Katz et al. 2015) that a MIC value, always calculated from longitudinal within-group data, can be applied between groups and used for calculation of the sample size needed to determine a between-group difference in mean PROM score. Applying responder analysis to KOOS data in orthopedic studies seems to solve this dilemma.

Bhandari M, Montori V M, Schemitsch E H. The undue influence of significant p-values on the perceived importance of study results. Acta Orthop 2005; 76 (3): 291-5.

Collins N J, Prinsen C A, Christensen R, Bartels E M, Terwee C B, Roos E M. Knee Injury and Osteoarthritis Outcome Score (KOOS): Systematic review and meta-analysis of measurement properties. Osteoarthritis Cartilage 2016; 24 (8): 1317-29.

Escobar A, Gonzalez M, Quintana J M, Vrotsou K, Bilbao A, Herrera-Espineira C, Garcia-Perez L, Aizpuru F, Sarasqueta C. Patient acceptable symptom state and OMERACT-OARSI set of responder criteria in joint replacement: Identification of cut-off values. Osteoarthritis Cartilage 2012; 20 (2): 87-92.

Felson D T, Anderson J J, Boers M, Bombardier C, Chernoff M, Fried B, Furst D, Goldsmith C, Kieszak S, Lightfoot R, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993; 36 (6): 729-40.

Katz N P, Paillard F C, Ekman E. Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. J Orthop Surg Res 2015; 10: 24.

King M T. A point of minimal important difference (MID): A critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2011; 11 (2): 171-84.

Ranstam J. P-values in research reports. Acta Orthop 2005; 76 (3): 289-90.

Ranstam J. Why the P-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage 2012; 20 (8): 805-8.

Roos E M. KOOS User's Guide. 2012. http://www.koos.nu/