Supplement 2:

Additional methods

In this Supplement we first explain our exact approach for calculating and defining the required effect sizes to be presented by the eight effect size indices. Then we also name further details regarding the presentation of the questions in our online questionnaire.

Following the recommendations by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT), we defined *response* for our scenario as a 50 % pain reduction and the MID as a 1-point VAS change (1). The pooled standard deviation (SD) of the change scores was assumed to be 2.5 (2, 3).

The effect sizes that were presented in our questionnaire were based on the widely accepted rule-of-thumb which defines a small effect as SMD=0.2, a medium effect as SMD=0.5 and a large effect as SMD=0.8 (4)

Using the SD=2.5, then, the MD corresponding with a small/medium/large effect was calculated as 0.5, 1.25, and 2.00 points, as changes in points on the VAS score between 0-10 respectively. To calculate the corresponding MID units, we used two previous publications as a guide, which defined a MID of 0.5 MID units as a small effect and a MID of 2 MID units as a large effect (5, 6). The mean value of these two represents the size of the MID for a medium effect (1.25 MID units). To calculate the values of RoM, an average endpoint score in pain after intervention had to be determined for our hypothetical scenario. After having reviewed the relevant literature, we agreed on an endpoint score of 5 points on the VAS on placebo (7, 8). Using this and the values of the MD, the values of the RoM could be calculated. Based on the SMD and Cohen's rule of thumb, we calculated the percentages corresponding with a small, medium or large effect using the formulae (9, 10). We used these values to calculate RR, RD, CER & EER, and NNT. The placebo response rate required for the calculations was set at 20% based on relevant literature and group consensus (11).

Since the presented effect size for every of the eight tested indices was randomized automatically in our questionnaire, we had to avoid that the random selection would favour one effect size disproportionally, e.g. a small effect size being shown eight times. We therefore introduced three different randomization combinations for the sizes of the presented effects:

1. 3x small effects, 3x medium effects, and 2x large effects
2. 3x small effects, 2x medium effects and 3x large effects
3. 2x small effects, 3x medium effects, and 3x large effects

We paid particular attention to the wording of the questions, as we believe that incorrect or inaccurate wording would have a negative impact on the understanding of the effect size measures. We also focused on the following aspects of the questions' design: The main questions should show a similar structure and verbalization had to be as comprehensible as possible. The treatment effects needed to be presented in the same way as they would appear in the results sections of medical publications, in order to be able to answer our research question.

References

1. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008;9(2):105-21; doi: 10.1016/j.jpain.2007.09.005.
2. Busse JW, Wang L, Kamaleldin M, et al. Opioids for Chronic Noncancer Pain: A Systematic Review and Meta-analysis. Jama. 2018;320(23):2448-60; doi: 10.1001/jama.2018.18472.
3. Enthoven WT, Roelofs PD, Deyo RA, et al. Non-steroidal anti-inflammatory drugs for chronic low back pain. Cochrane Database Syst Rev. 2016;2(2):Cd012087; doi: 10.1002/14651858.Cd012087.
4. Cohen J. Statistical power analysis for the behavioral sciences. 2. ed. Hillsdale, NJ [u.a.]: Erlbaum; 1988.
5. Johnston BC, Alonso-Coello P, Friedrich JO, et al. Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. Cmaj. 2016;188(1):25-32; doi: 10.1503/cmaj.150430.
6. Johnston BC, Thorlund K, Schünemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. Health Qual Life Outcomes. 2010;8:116; doi: 10.1186/1477-7525-8-116.
7. Allegrini A, Nuzzo L, Pavone D, et al. Efficacy and safety of piroxicam patch versus piroxicam cream in patients with lumbar osteoarthritis. A randomized, placebo-controlled study. Arzneimittelforschung. 2009;59(8):403-9; doi: 10.1055/s-0031-1296415.
8. Kivitz AJ, Gimbel JS, Bramson C, et al. Efficacy and safety of tanezumab versus naproxen in the treatment of chronic low back pain. Pain. 2013;154(7):1009-21; doi: 10.1016/j.pain.2013.03.006.
9. Furukawa TA. From effect size into number needed to treat. Lancet. 1999;353(9165):1680; doi: 10.1016/s0140-6736(99)01163-0.
10. Furukawa TA, Leucht S. How to Obtain NNT from Cohen's d: Comparison of Two Methods. PLOS ONE. 2011;6(4):e19070; doi: 10.1371/journal.pone.0019070.
11. Moore A, Derry S, Eccleston C, Kalso E. Expect analgesic failure; pursue analgesic success. Bmj. 2013;346:f2690; doi: 10.1136/bmj.f2690.