Journal of Clinical Epidemiology 191 (2026) 112114

ORIGINAL RESEARCH

The interplay between PROM score distributions and treatment effect detection likelihood in randomized controlled trials—a metaepidemiologic study

Valtteri Panula[a,*], Antti Saarinen[b], Matias Vaajala[c], Rasmus Liukkonen[c], Oskari Pakarinen[d], Juho Laaksonen[c], Ville Ponkilainen[a], Ilari Kuitunen[e,f], Mikko Uimonen[c,g]

[a]*Center for Musculoskeletal Diseases, Tampere University Hospital, Tampere University, Tampere, Finland*
[b]*Department of Orthopaedics and Traumatology, Turku University Hospital, University of Turku, Turku, Finland*
[c]*Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*
[d]*Department of Surgery, Päijät-Häme Central Hospital, Lahti, Finland*
[e]*Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland*
[f]*Department of Pediatrics, Kuopio University Hospital, Kuopio, Finland*
[g]*Heart Hospital, Tampere University Hospital, Wellbeing Services County of Pirkanmaa, Tampere, Finland*

Accepted 15 December 2025; Published online 19 December 2025

Abstract

Objectives: We hypothesized that, in musculoskeletal randomized controlled trials (RCTs) using patient-reported outcome measures (PROMs), higher baseline scores and the clustering of follow-up scores near the upper bound (ie, ceiling effect) compress variability and attenuate measurable between-group differences, thereby lowering the likelihood of observing a statistically significant effect. We therefore examined how score distributions at pretreatment and follow-up influence the likelihood of detecting between-group differences.

Study Design and Setting: We conducted a metaepidemiologic study of RCTs, published between 2015 and 2024, that compared treatment effects on musculoskeletal disorders between two study groups using PROMs. The observed distributions of the PROM scores at baseline and follow-up were collected from the included studies. All PROM scores were rescaled to 0-100 with higher scores indicating better health. The likelihood of observing a statistically significant difference in PROM scores between the study groups was examined by calculating the score difference required to achieve a P value $<.05$.

Results: A total of 255 RCTs were included. PROM scores improved from baseline to follow-up in most studies (98%), with a mean change of +28 points. The correlation coefficient between the mean baseline score and mean score change was −0.66 (95% CI -0.72 to −0.59) indicating that higher baseline scores were associated with lower score change. In addition, there was a moderate correlation between the mean and SD of PROM scores at follow-up (−0.39; 95% CI -0.48 to −0.28). The mean likelihood of detecting a difference was 65% (SD 11%) at baseline and 65% (SD 11%) at follow-up. The likelihood reached the 80% benchmark in only 8.5% and 8.1% of the studies at baseline and follow-up, respectively.

Conclusion: The concentration of PROM score distributions toward the high end of the scale, especially when higher baseline scores are present, diminishes the likelihood of detecting significant differences between study groups, particularly at follow-up assessments in studies analyzing musculoskeletal complaints. This underscores the importance of critically evaluating the conclusions drawn from these studies.

Keywords: Metaepidemiology; Musculoskeletal diseases; Patient-reported outcome measure; PROM; COSMIN; RCT

1. Introduction

Randomized controlled trials (RCTs) represent the highest level of evidence in clinical treatment research and carry substantial weight in the development of treatment guidelines [1]. The widespread adoption of patient-reported

Funding: This research did not receive any specific grant from funding agency in the public, commercial, or not-for-profit sectors.

* Corresponding author. Center for Musculoskeletal Diseases, Tampere University Hospital, Tampere University, Elämänaukio 2, FI-33520, Tampere, Finland.
E-mail address: valtteri.panula@tuni.fi (V. Panula).

https://doi.org/10.1016/j.jclinepi.2025.112114

What is new?

Key findings

- Across published musculoskeletal trials, we found that patient-reported outcome measures (PROMs) often do not provide a fair chance to detect true differences between treatment groups. This is mainly because many patients already report high scores at the start of the study, leaving little room for improvement to be captured by the scale.

What this adds to what is known?

- Our study shows that when PROM scores cluster near the top of the scale, particularly at follow-up, even meaningful differences between treatments can remain hidden. This limitation is built into the way these scales work and affects how reliably they can detect change.

What is the implication and what should change now?

- Because PROMs may not always be sensitive enough to show differences between groups, ''no difference'' results should be interpreted with caution. Researchers can use information available at baseline to anticipate when PROM scales are likely to miss important changes and adjust study designs accordingly, ensuring that outcomes better reflect patients' true clinical status.

outcome measures (PROMs) as tools for outcome assessment has enabled the evaluation of patients' subjective experiences when assessing treatment efficacy. Since the adoption of PROMs in clinical research, attention has increasingly focused on ineffective treatments and their removal from clinical guidelines [2]. The basis for discontinuing such treatments has often been RCTs that, using PROMs, demonstrate no advantage for a treatment over a comparator or placebo [3,4].

However, the use of PROMs in clinical trials carries a significant risk of ceiling effect, a situation in which scores cluster at the upper limit of a measurement scale, limiting the ability to detect improvement or differences. Ceiling effects arise from a mismatch between the PROM scale and the clinical states and expected outcomes of the patient sample, as well as from the psychometric properties of the instrument which is common when PROMs that are inadequately developed (= the majority of existing PROMs) are used in RCTs [5−7]. They reduce the discriminative power of PROMs as scores accumulate toward the upper end of the scale, which can lead to a heightened risk of erroneous conclusions, suggesting no difference between treatments, even when substantial differences in efficacy may exist [5]. Ceiling effects can result from using PROMs originally developed for severely symptomatic conditions (eg, advanced osteoarthritis) in less symptomatic patient groups, from measurement scales that lack invariance (especially if not Rasch-validated), or from scoring systems that combine multiple dimensions into a single score, which can dilute changes in individual domains and thus reduce the measure's ability to detect improvement.

When using PROMs to evaluate treatment response and compare scores between groups, it is possible, under certain conditions, to estimate the likelihood of detecting a difference (ie, score difference or change that is considered as relevant) even before treatment initiation and follow-up. This likelihood can be determined based on patients' baseline clinical status. If it is assumed that clinical status in both groups will not worsen after follow-up and the PROM scores will not decrease, then the likelihood of detecting a difference depends on the average baseline PROM score of patients and the number of points required for this difference. If we further assume that follow-up scores are determined randomly, the likelihood of detecting the difference due to chance can be calculated using the following formula, which is based on our theoretical considerations:

$$P(\textit{Required difference}) = \left(\frac{100 - \textit{Baseline score} - \textit{Required difference}}{100 - \textit{Baseline score}}\right)^2$$

This likelihood is composed of a double conditional probability: [1] that the treatment group score exceeds the comparator group score, and [2] that the treatment group score meets or exceeds the target point difference required to achieve the clinically or statistically significant difference compared to the comparator group. By definition of the formula, higher observed PROM scores and larger required score differences are associated with a lower likelihood of detecting the difference between the groups.

In the context of an RCT, at the time of randomization, the groups can be assumed to be comparable in terms of pretreatment clinical state and PROM score distributions. Therefore, given the presented formula, the likelihood of observing a significant difference can initially be estimated using the mean score of the population from which the study sample is drawn, that is, assumed baseline score. At follow-up, however, this likelihood is determined primarily

by the mean score of the comparator group, which serves as a reference score for the likelihood calculation and is comparable to the baseline score defined in the formula.

The aim of this metaepidemiologic study was to examine how score variability at the pretreatment and follow-up phases influences the likelihood of detecting a difference between treatment groups in RCTs using PROMs as outcome measures in patients with musculoskeletal complaints. We hypothesized that higher baseline scores are associated with a smaller-magnitude change due to proximity to the upper boundary of the PROMs' measurement scale and that accumulation of PROM scores toward the upper end of the scale would result in lower variability. Furthermore, we anticipated that the likelihood of detecting a difference is related to the observed PROM score distribution (group means, SDs, and sample sizes per arm).

2. Methods

2.1. Search process

This metaepidemiologic study was reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Meta-Epidemiological extension guidelines [8]. The search was conducted in PubMed (National Library of Medicine) and Web of Science (Clarivate) on March 15, 2024. The search was limited to studies published between 2015 and the search date to obtain the most recently published studies. The search strategy for both databases is presented in the Supplementary materials.

2.2. Inclusion and exclusion criteria

All RCTs comparing two patient groups with musculoskeletal complaints using a PROM as an outcome measure were included. Studies were excluded if they met at least one of the following criteria: (i) included patients with traumatic injuries (defined as musculoskeletal injuries caused by sudden physical trauma, eg, falls, collisions, or high-energy impact); (ii) compared more than two groups; or (iii) used only a single-item visual analog scale as the outcome measure, with no additional PROMs included.

2.3. Review process

After the initial search, the results were uploaded to Covidence (Veritas Healthcare Inc, Melbourne, Australia) software for screening [9]. The screening was conducted in duplicate as two authors independently screened titles and abstracts. In case of disagreement between the two reviewers, consensus was achieved through the opinion of a third author. Full texts were independently screened by two authors to assess the eligibility of the data. The extraction was piloted (10% of the included studies) by two extraction authors and once sufficient consensus was achieved (weighted kappa >0.80); single extraction was performed.

During the extraction, the following information was collected from the included studies: name of the journals, name of the used PROMs, scales of the PROMs, sample sizes, PROM score distribution parameters (means, SDs, medians, IQRs, 95% CIs, SEs), statistical tests used, P values, and possible reported differences on treatment effects between groups based on the PROM scores. Articles with missing information were excluded. The PROM scores were collected from the baseline and from the latest follow-up time point.

2.4. Statistical analysis

PROM scores were standardized to a 0−100 scale, with higher scores indicating better outcomes. When only a 95% CI or SE was provided, SD was calculated accordingly. If only median and IQR were reported, the mean and SD were estimated using a rough conversion algorithm: the conversion of medians to means was performed by substituting the median with the mean, whereas for IQRs, the SD was approximated using the formula

$$SD = \frac{(Upper\ bound\ of\ IQR - Lower\ bound\ of\ IQR)}{1.35}$$

This formula was based on the properties of the normal distribution, where the IQR corresponds to 1.35 times the SD. The IQR represents the middle 50% of the data, and in a normal distribution, its relationship to the SD is constant. Even when approximated from nonnormal data, the IQR-based estimate was considered sufficiently accurate as strict distributional assumptions were not the focus in the analysis.

To test the correctness of the assumption that clinical state of the patients does not worsen after the follow-up and thereby the PROM scores do not decrease, the PROM score distributions were assessed at baseline and follow-up. The proportion of studies in which the PROM scores improved from baseline to follow-up was calculated. Furthermore, the strength of the association between the baseline PROM score and score change from baseline to follow-up was calculated using Pearson's correlation coefficient. Similarly, Pearson's correlation coefficient was calculated between the mean and SD of PROM scores at follow-up.

The score difference required for statistical significance ($P < .05$) between groups, a critical point difference, was then calculated using the following formula derived from the standard two-sample z-test for comparing means:

$$Critical\ point\ difference = \left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right) \times \sqrt{\left(\frac{SD_1{}^2}{n_1}\right) + \left(\frac{SD_2{}^2}{n_2}\right)}$$

In this formula, $Z_{\alpha/2}$ represent the Z value from a normal distribution with significance level $\alpha = 0.05$ giving $Z_{0.025} \approx 1.96$ and Z_{β} represent the Z value of the statistical power $\beta = 1 -$ statistical power $= 1 - 0.8$ giving $Z_{0.2} \approx 0.84$. SD_1 and SD_2 represent the SDs, and n_1 and n_2 represent the sample sizes of groups 1 and 2, respectively.

To evaluate retrospectively each trial's statistical power, that is, the likelihood of detecting a difference due to chance, a simulation-based analysis method was used. The likelihood represents the study's ability to detect a statistically significant difference between study groups if the difference exists in population. For each study, two normal distributions were simulated, representing the scores of two treatment groups at baseline and follow-up. The parameters of the distributions—mean, SD, and sample size—were defined based on the source data for each study. In each iteration, the mean differences, that is, the differences between group means ($\Delta = \overline{X}_1 - \overline{X}_2$) were calculated from the simulated distributions. The likelihood of detecting a difference was calculated from the simulated mean differences as follows:

$$Likelihood = P(|\Delta| \geq Critical\ point\ difference)$$

where $|\Delta|$ represents the simulated mean differences and the critical point difference is the study-specific critical threshold. Each study underwent 1000 simulations to ensure a sufficiently accurate probability estimate and the likelihood is calculated by the proportion of the simulations in which $|\Delta|$ was higher than the critical point difference. If a study's likelihood is below 50%, such study has less than a 50% chance of detecting an existing difference. In other words, its ability to detect a difference is worse than a random guess. A low likelihood ($<50\%$) indicates a study's weak ability to detect a difference, which may result from small sample sizes or large SDs. Uniformly with the generally used reference for sufficient statistical power (type I error) when designing RCT and calculating sample sizes, a likelihood $>80\%$ was interpreted as that the study was statistically robust in detecting significant differences. The statistical analysis was conducted using R (version 4.4.2; R Foundation for Statistical Computing).

3. Results

The initial search identified 5451 studies. After abstract screening, 665 full texts were assessed of which 255 fulfilled the inclusion criteria and were included in the analysis. In these studies, the median sample size was 79.5 patients (IQR 50−124.5). In all, 81 (31%) of the studies reported observing a statistically significant difference between the study groups at the end of the follow-up.

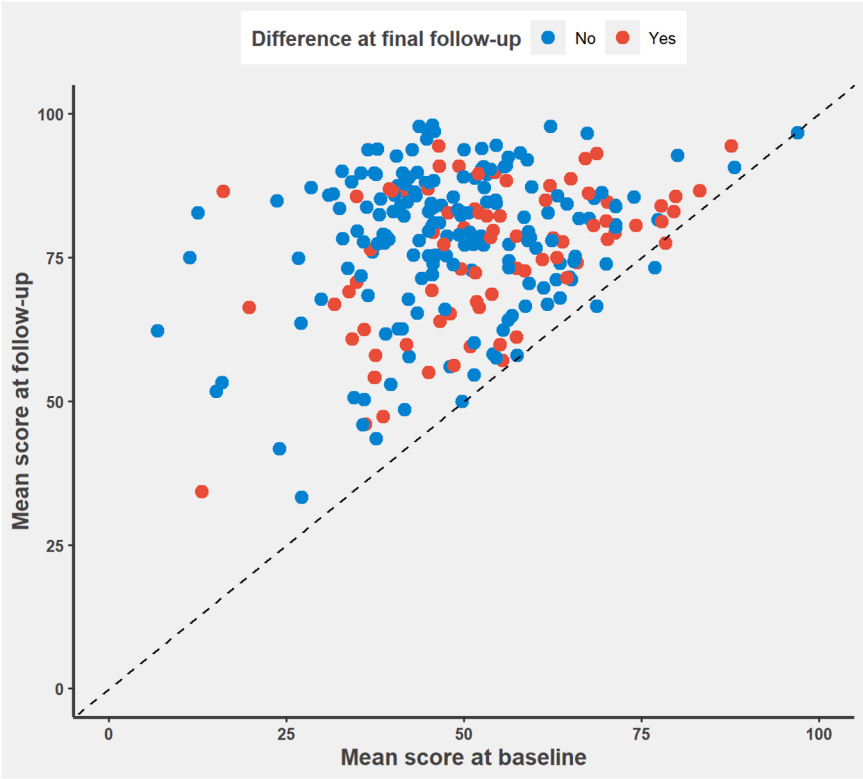

Figure 1. Distribution of the study-wise mean baseline scores in relation to the mean follow-up scores. Upper half above the dashed line indicates that the mean follow-up score was higher than the mean baseline score. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

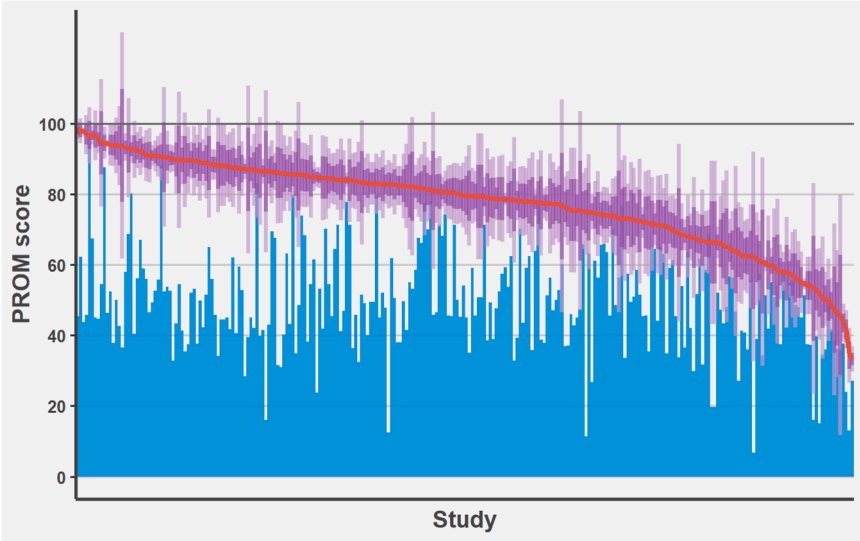

Figure 2. The distributions of study-wise mean baseline scores (light blue) and mean follow-up scores (red). Purple area represents the mean follow-up score ± half of the critical point difference and light purple mean follow-up score ± the critical point difference. In studies with purple bar exceeds 100, it is not possible to observe a statistically significant difference and in studies with light purple bar exceeds 100, the sensitivity to observe a statistically significant difference is decreased. PROM, patient-reported outcome measure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

PROMs were defined as the primary endpoint in 179 studies and as the secondary endpoint in 76 studies.

The mean PROM score was 50 (SD 14) at the baseline and 77 (SD 13) at the follow-up. The mean change from baseline to follow-up was +28 (SD 16). In 98% of the studies, the mean PROM scores improved from the baseline to the follow-up (Fig 1). Pearson correlation coefficient between the mean baseline score and mean score change was −0.66 (95% CI −0.72 to −0.59) indicating that lower baseline score were associated with higher score change. Furthermore, Pearson correlation coefficient between the mean and SD of the PROM scores at the follow-up was −0.39 (95% CI −0.48 to −0.28).

The mean critical point difference at follow-up was 11 (SD 6.0). In six studies (2.3%), adding a half of the critical point difference to the mean follow-up score resulted in scores exceeding 100 and thereby out of the measurement scales (Fig 2). In studies with such follow-up score mean and critical point difference, it is not even possible to observe a statistically significant difference. Furthermore, in 28 studies (11%), adding the critical point difference to the mean follow-up score resulted in scores exceeding 100. In such studies, the sensitivity to observe a statistically significant difference is decreased. When adding the expected score change, that is, the overall mean change in PROM scores from baseline to follow-up, to the observed baseline scores, the expected follow-up score exceeded 100 points in 14 (5.4%) studies, 95 points in 32 (12%) studies, and 90 points in 51 (20%) studies.

The examination of the likelihood of detecting a difference showed generally decreased likelihood estimates as at the baseline only 22 (8.5%) studies and at the follow-up only 21 (8.1%) studies exceeded 80%. The mean likelihood was 65% (SD 11%) at the baseline and 65% (SD 11%) at the follow-up. At both baseline and follow-up, 23 (8.8%) studies had likelihood lower than 50% indicating that the ability to detect a true difference was even weaker than a random guess. When comparing the studies that did report a statistically significant difference between the study groups with those that did not, the difference was rather modest (66%, SD 13% in studies reporting a difference vs 65%, SD 11% in studies reporting no difference).

4. Discussion

The aim of this metaepidemiologic study was to demonstrate how the likelihood of detecting a statistically significant difference in PROM-based outcomes in musculoskeletal RCTs is related to the PROM score distribution. As our results show, it is not rare that the accumulation of PROM scores toward the high end of the scale decreases the possibility of observing a significant difference between the study groups. Thus, it is essential to give critical consideration to the conclusions that can be drawn in such studies.

As shown by numerous, often cited trials using PROMs as an outcome measure, extending the follow-up period is associated with the accumulation of scores toward high end of the PROM scale increasing the risk for ceiling effect [3,4,10−13]. Indeed, a recent study examining effect sizes of elective orthopedic surgery using various PROMs in various musculoskeletal conditions demonstrated a notable accumulation of PROM scores toward the high end of the

scale even after a 1-year follow-up, regardless of the PROM used or condition treated [6]. In light of these studies, it is reasonable to assume that the ceiling effect will become even more apparent as the follow-up period is prolonged.

A significant ceiling effect diminishes the statistical power to detect a difference between patient groups, leading to potentially erroneous conclusions of no difference in the treatment effect [5]. This limitation reflects both the mathematical properties of bounded measurement scales and the psychometric behavior of ordinal instruments. Mathematically, a PROM with a fixed upper limit imposes a truncation on the distribution: as group means approach the ceiling, the remaining observable range diminishes, variance becomes compressed, and latent improvements cannot be expressed once the boundary is reached. Under these conditions, the observed between-group difference must necessarily underestimate the true difference, and the SE becomes large relative to the constrained effect size, resulting in a systematic loss of statistical power. Psychometrically, most PROMs are ordinal and inherently nonlinear, meaning that identical point changes reflect different amounts of clinical change depending on where they occur on the scale. It has been shown that even PROMs designed to approximate interval-level measurement become increasingly nonlinear near the upper end, where item responses saturate and the instrument's ability to discriminate between high-functioning individuals collapses [14]. In this region, response categories no longer separate clinically meaningful gradations in health status, and the minimum detectable difference is strongly dependent on the starting position on the scale. Importantly, these phenomena are not simply artifacts of poorly constructed PROMs but arise from structural constraints of bounded ordinal measurement systems. Ceiling effects therefore represent an inherent limitation that restricts observable differences between groups, even when genuine differences in latent health states are present.

In 98% of the studies in this review, the PROM scores improved from baseline to follow-up, and the mean score change was +28. With regard to these findings, it is reasonable to consider the PROM score observed at the beginning of the study period as a ground level for the expectable PROM scores, from which the score may be assumed to improve with reasonable certainty. Therefore, if clinical researchers observe a high baseline mean PROM score, they should acknowledge that the probability of observing a statistically significant difference at follow-up is low or even nonexistent. Indeed, such an association has already been demonstrated in knee arthroplasty patients [15]. Furthermore, the mean score difference needed to a statistically significant difference at the follow-up was 11 points. As a result, with a mean increase of 28 points and 11 points needed for a statistically significant difference, and thereby approximately a 33.5-point increase in one group and a 22.5-point increase in the other, it seems highly unlikely to observe a significant difference at follow-up if the baseline score exceeds 70. In this review, the proportion of studies in which the mean baseline score exceeded 70 was 8.5%, exceeded 65 was 14% and exceeded 60 was 22%. With regard to the previously mentioned study on the effect sizes of elective orthopedic surgeries, the minimal clinically important difference values for the PROMs may be assumed to fall between 6.0 and 10.5 points, depending on the PROM used and condition treated [6]. Hence, if the follow-up score exceeds 90, the justification for concluding no difference between the patient groups is questionable. According to this review, in 13% of the studies, the mean follow-up score was 90 or higher, and in these studies, the mean difference between the groups was 1.1 points. Retrospectively, adding the mean PROM score change calculated in this review, that is, +28, to each study's mean baseline score, 20% of the studies would have been expected to exceed 90 points. Although the negative correlation between the baseline score and score change indicates that such retrospective estimation of the expected follow-up score may overestimate the expected follow-up score, it should be acknowledged that the risk of false-negative conclusions due to the ceiling effect increases with increasing baseline score. In addition, the magnitude of this risk may be estimated by using the baseline scores, regardless of certain shortcomings. The value of risk assessment beforehand lies in enabling the redesign of the outcome measurement instruments to better cover the patients' expected clinical state at follow-up. Lastly, the likelihood of detecting a difference reached the benchmark of 80% only in 8.1% of the studies, with the mean likelihood of 65%. On the other end of the spectrum, in 8.8% of the studies the likelihood was less than 50% representing a lower ability to detect an existing difference than a random guess. Overall, these findings highlight the increasing risk of unjustified conclusions of no difference between study groups as scores accumulate toward the boundaries of the PROM's measurement scale.

From a broader perspective, the results of this review underscore that a statistically nonsignificant difference in PROM scores does not exclude a clinically important difference between study groups if the scores have accumulated toward the high end of the PROM scale. It is the responsibility of authors to make justified conclusions regarding the findings their research yields. With respect to PROM scores, and thereby the measurement spectrum of the PROM, a nonsignificant difference in a presence of a ceiling effect should be interpreted to mean that the groups are indeed equal in light of the measurement spectrum of the given PROM. However, more importantly, PROM scores presented with a significant ceiling effect do not genuinely represent the true clinical state of the patients, which may be beyond the PROM scale. As an example, consider a situation in which a certain hip arthroplasty prosthesis reduces symptoms related to hip joint osteoarthritis to the point where no symptoms remain. Intuitively, the outcomes using such a prosthesis may be

considered unsurpassable. Now, a novel hip implant not only reduces symptoms but also enables patients to return to sports, for example, play football or begin running. If the previous implant has not enabled such functionality but only reduced symptoms to nonexistence, the novel implant is undoubtedly better than the previous one. If a PROM does not capture the ability to return to sports after hip arthroplasty but only symptoms of osteoarthritis, then the aforementioned implants would seem similarly good despite the clear difference. Thus, by selecting the PROM to be used in outcome assessment, clinical researchers themselves set the boundaries for the expectable improvement. Indeed, it is not uncommon for RCTs examining patients with musculoskeletal complaints to use inadequate PROMs with deficient content validity in outcome measurement, which in turn has been shown to reduce the likelihood of detecting a significant difference between patient groups [16]. As an example of such a situation, a prior RCT by Frobell et al [17] used four dimensions of the Knee injury and Osteoarthritis Outcome Score (KOOS) as a primary outcome for patients with anterior cruciate ligament (ACL) injury. Although KOOS has low content validity in ACL patients and therefore the validity of the study's findings is questionable, its findings showing no difference between the treatments have nevertheless had a substantial influence on clinical practice guidelines. To avoid erroneous conclusions and misleading recommendations, interpretation of such results should be humble and accurate.

Notwithstanding the foregoing, accurate and honest interpretations should not be considered the sole solution to the problem. It is obvious that a clinical trial using an outcome measure that does not capture the clinical state of the patient at the time of deciding which treatment is more effective is not of value for clinical decision-making. A remedy would be to apply outcome measures sensitive enough to detect nuances of the clinical state at the time of the decision, that is, PROMs of which validity and responsiveness have been established in a patient population after an adequate follow-up time. However, the development and validation of PROMs are targeted to a relatively short time period during or after the acute phase of a musculoskeletal health complaint, leading to worse measurement performance after long time periods. Another solution would be to adopt other relevant outcome measures, such as return to work, duration of sick leave, and need for pain medication. Furthermore, the probability of observing a statistically significant difference at follow-up may be estimated using the baseline scores, which enables the redesigning of the study protocol to better cover the patients' expected clinical state at follow-up and thereby improve the sensitivity to detect potential differences between patient groups. The authors encourage the development of alternative outcome measures that are not as dependent on limited measurement scale or confounding from psychological factors and habituation after long follow-up periods.

4.1. Strengths and limitations

This review had several limitations. First, the conceptualization of this review is theory-based, and thus the findings should be interpreted cautiously. The estimation of the likelihood of observing a difference relied completely on the assumption that the follow-up score is determined by chance and randomness without other dependencies, which does not comply with the real-world setting. In addition, estimating the score difference needed to achieve a statistically significant difference may hold computational nuances, causing uncertainty in the estimates. However, our approach is illustrative of the issue under investigation and thus we found it justified, as the results do not translate straight to clinical practice but rather to research methodology without a strict need for accurate effect estimates. Secondly, the decision to convert the medians and IQRs reported in the study papers to means and SDs may not be justified from a statistical perspective. However, our intention was to demonstrate the problem related to the ceiling effect, and we presumed that in the studies reporting medians and IQR, that is, studies observing a skewed PROM score distribution, the ceiling effect would be even more profound and thereby the problem even more severe, which was regarded as a strong incentive to perform the conversion and to include these studies in the analysis despite the relatively small inaccuracy from forcing the distributional conversion. Third, we included only RCTs comparing two patient groups involving patients with musculoskeletal complaints, although the issues related to a limited measurement scale and ceiling effect are related to all PROMs regardless of the conditions to be treated and the number of patient groups. Fourth, as we did not have access to patient- and item-level data, we could not further assess the probability of observing a statistically significant difference from a psychometric point-of-view, involving an assessment of patient-item distribution across the PROM scales. Fifth, although the search strategy may have excluded some potentially eligible studies, this does not undermine the demonstrative nature of the present study. Lastly, this study was not preregistered in any public database before its conduct, which may introduce a risk of selective reporting and analytical flexibility. The decision not to preregister the protocol was made based on the exploratory and demonstrative nature of our analytical approach, which we anticipated would require unavoidable amendments in the analysis plan. The risk was counterbalanced by adhering to established high-quality methodological standards in analytics and by reporting the methodology transparently.

5. Conclusion

The concentration of PROM score distribution toward the high end of the scale, especially when higher baseline

scores are present, diminishes the likelihood of detecting significant differences between study groups, particularly at follow-up assessments in studies analyzing musculoskeletal complaints. This underscores the importance of critically evaluating the conclusions drawn from these studies.

CRediT authorship contribution statement

Valtteri Panula: Writing − original draft, Project administration, Methodology, Data curation, Conceptualization. **Antti Saarinen:** Writing − review & editing, Software, Project administration, Methodology, Data curation, Conceptualization. **Matias Vaajala:** Writing − review & editing, Visualization, Validation. **Rasmus Liukkonen:** Writing − review & editing, Visualization, Validation. **Oskari Pakarinen:** Writing − review & editing, Visualization, Validation. **Juho Laaksonen:** Writing − review & editing, Visualization, Validation. **Ville Ponkilainen:** Writing − review & editing, Visualization, Validation. **Ilari Kuitunen:** Writing − review & editing, Visualization, Validation. **Mikko Uimonen:** Writing − review & editing, Visualization, Supervision, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

A.S. received financial support from Vappu Uuspää Foundation, and Päivikki and Sakari Sohlberg Foundation. There are no competing interests for any other author.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.112114.

Data availability

Data will be made available on request.

References

[1] Wallace SS, Barak G, Truong G, Parker MW. Hierarchy of evidence within the medical literature. Hosp Pediatr 2022;12(8):745−9.
[2] Kluzek S, Dean B, Wartolowska KA. Patient-reported outcome measures (PROMs) as proof of treatment efficacy. BMJ Evid Based Med 2022;27(3):153−5.
[3] Paavola M, Malmivaara A, Taimela S, Kanto K, Inkinen J, Kalske J, et al. Subacromial decompression versus diagnostic arthroscopy for shoulder impingement: randomised, placebo surgery controlled clinical trial. BMJ 2018;362:k2860.
[4] Sihvonen R, Paavola M, Malmivaara A, Itälä A, Joukainen A, Kalske J, et al. Arthroscopic partial meniscectomy for a degenerative meniscus tear: a 5 year follow-up of the placebo-surgery controlled FIDELITY (Finnish Degenerative Meniscus Lesion Study) trial. Br J Sports Med 2020;54:1332.
[5] Saarinen A, Pakarinen O, Vaajala M, Liukkonen R, Ponkilainen V, Kuitunen I, et al. Randomized controlled trials reporting patient-reported outcomes with no significant differences between study groups are potentially susceptible to unjustified conclusions—a systematic review. J Clin Epidemiol 2024;169:111308.
[6] Äärimaa V, Kohtala K, Rantalaiho I, Ekman E, Mäkelä K, Taskinen HS, et al. A comprehensive approach to PROMs in elective orthopedic surgery: comparing effect sizes across patient subgroups. J Clin Med 2024;13(11):3073.
[7] Blackburn M, Kaplan SL. Are priorities of younger patients with knee pain addressed by PROMs? A qualitative study. Phys Ther Sport 2019;40:160−8.
[8] Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139−42.
[9] Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia [Internet]. Available at https://www.covidence.org/. Accessed January 11, 2025.
[10] Beard DJ, Rees JL, Cook JA, Rombach I, Cooper C, Merritt N, et al. Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. Lancet 2018;391:329.
[11] Ziga M, Sosnova M, Zeitlberger AM, Regli L, Bozinov O, Weyerbrock A, et al. Objective outcome measures may demonstrate continued change in functional recovery in patients with ceiling effects of subjective patient-reported outcome measures after surgery for lumbar degenerative disorders. Spine J 2023;23(9):1314−22.
[12] Eckhard L, Munir S, Wood D, Talbot S, Brighton R, Walter B, et al. The ceiling effects of patient reported outcome measures for total knee arthroplasty. Orthop Traumatol Surg Res 2021;107(3):102758.
[13] Paavola M, Kanto K, Ranstam J, Malmivaara A, Inkinen J, Kalske J, et al. Subacromial decompression versus diagnostic arthroscopy for shoulder impingement: a 5-year follow-up of a randomised, placebo surgery controlled clinical trial. Br J Sports Med 2021;55:99−107.
[14] Spratt KF. Minimal clinically important difference based on clinical judgment and minimally detectable measurement difference: a rationale for the SF-36 physical function scale in the SPORT intervertebral disc herniation cohort. Spine (Phila Pa 1976) 2009;34(16):1722.
[15] Clement ND, Afzal I, Demetriou C, Deehan DJ, Field RE, Kader DF. The preoperative Oxford Knee Score is an independent predictor of achieving a postoperative ceiling score after total knee arthroplasty. Bone Joint J 2020;102-B(11):1519−26.
[16] Hansen CF, Jensen J, Brodersen J, Siersma V, Comins JD, Krogsgaard MR. Are adequate PROMs used as outcomes in randomized controlled trials? An analysis of 54 trials. Scand J Med Sci Sports 2021;31(5):972−81.
[17] Frobell RB, Roos EM, Roos HP, Ranstam J, Lohmander LS. A randomized trial of treatment for acute anterior cruciate ligament tears. N Engl J Med 2010;363:331−42.