Journal of Clinical Epidemiology 169 (2024) 111308

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Randomized controlled trials reporting patient-reported outcomes with no significant differences between study groups are potentially susceptible to unjustified conclusions—a systematic review

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

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

Accepted 21 February 2024; Published online 28 February 2024

Abstract

Objectives: Ceiling effect may lead to misleading conclusions when using patient-reported outcome measure (PROM) scores as an outcome. The aim of this study was to investigate the potential source of ceiling effect−related errors in randomized controlled trials (RCTs) reporting no differences in PROM scores between study groups.

Study Design and Setting: A systematic review of RCTs published in the top 10 orthopedic journals according to their impact factors was conducted, focusing on studies that reported no significant differences in outcomes between two study groups. All studies published during 2012−2022 that reported no differences in PROM outcomes and used parametric statistical approach were included. The aim was to investigate the potential source of ceiling effect−related errors—that is, when the ceiling effect suppresses the possible difference between the groups. The proportions of patients exceeding the PROM scales were simulated using the observed dispersion parameters based on the assumed normal distribution, and the differences in the proportions between the study groups were subsequently analyzed.

Results: After an initial screening of 2343 studies, 190 studies were included. The central 95% theoretical distribution of the scores exceeded the PROM scales in 140 (74%) of these studies. In 33 (17%) studies, the simulated patient proportions exceeding the scales indicated potential differences between the compared groups.

Conclusion: It is common to have a mismatch between the chosen PROM instrument and the population being studied increasing the risk of an unjustified ''no difference'' conclusion due to a ceiling effect. Thus, a considerable ceiling effect should be considered a potential source of error.

Keywords: Patient-reported outcome measure; PROM; Validity; Validation study; Methodology; Ceiling effect; COSMIN

1. Introduction

Patient-reported outcome measures (PROMs) are commonly used to evaluate treatment efficacy, track disease progression, and monitor patient outcomes in orthopedic science [1]. PROMs are standardized questionnaires designed to assess patients' subjective experiences of their health and the impact of interventions on their quality of life. The use of PROMs has led to a better understanding of the key aspects of an effective treatment from the patients' perspectives, thereby improving outcomes and enhancing mutual understanding between patients and healthcare providers. However, the ability of a PROM to reflect a patient's clinical condition depends on its psychometric properties [2]. Problems with these properties may result in inaccurate reflections of patients' clinical states,

* Corresponding author. Department of Orthopaedics and Traumatology, Turku University Hospital, Finland, University of Turku, Clinical Department, Orthopedics and Traumatology, PO Box 28, 20701, Turku, Finland.
E-mail address: anjusaa@utu.fi (A. Saarinen).

https://doi.org/10.1016/j.jclinepi.2024.111308

What is new?

- In this systematic review, out of 190 included studies, 74% exhibited scores exceeding the PROM scales, indicating the presence of a ceiling effect.
- A ceiling effect can prevent the detection of true differences between study groups, leading to unjustified conclusions.
- Ceiling effect should be considered as a potential source of error when comparing PROM scores between groups.

which can lead to erroneous conclusions concerning exposures and outcomes, thereby negatively affecting clinical decision-making and patient care [3].

Psychometric evaluation of PROMs is based on two theoretical paradigms: classical test theory and item-response theory [4]. PROMs aim to capture the implicitly continuous nature of clinical states and represent them on a continuous linear scale. The main idea is to transform the otherwise intangible and subjective clinical state of a patient into a numerical score, which is believed to reflect their clinical condition on a specific linear scale. Practically, this transformation is achieved by asking the patient a series of simple and detailed questions that cover different aspects of their clinical state. Each answer is assigned a score, and the sum or other compound value of all the question scores is assumed to provide a comprehensive representation of the patient's actual clinical state. Thus, the final score is considered an all-encompassing measure of the patient's clinical condition.

One of the main limitations of PROMs is that they can only measure a limited range of the entire spectrum of clinical states with only a limited perspective to its different domains. This is because each question in the PROM only captures a specific aspect of the clinical state with a limited scale and it is not possible to include an infinite number of questions with an infinite number of response categories in a PROM to fully capture the complexity of a patient's clinical state. Measuring outcomes using a PROM sets boundaries to the measurable spectrum of clinical states as well as the clinical state domains to be measured. Therefore, the choice of a PROM in use includes the selection of the specific yet limited domain of clinical states to be measured, as well as the spectrum of observable outcomes within the given domain, while disregarding other domains and clinical states outside the measurement scale.

A PROM is developed and validated in a certain patient group to reflect the outcomes in that group, whereas it may not perform equally well among patients with different clinical states. When the PROM does not fit the qualities of a given patient group, patients' clinical states may not be covered by the PROM's measurement scale, which leads to extreme scores. The concentration of scores toward the high end of a scale is referred to as the ceiling effect, whereas their concentration toward the low end of the scale is known as the floor effect [5]. It is important to understand that the limits imposed by a PROM scale do not necessarily represent the true limits of a patient's ability, quality of life, or other aspects of their clinical state in real life. For example, being completely free from disability, which is often indicated by a maximum PROM score, does not equal achieving the best possible ability. Improvements in a patient's real-life ability may go beyond the artificial boundaries set by the PROM's scale, which only reflects the absence of disability, leading to a ceiling effect in the PROM scores.

In trials comparing the effects of exposures on patient outcomes, it is expected that the patients' clinical states will change over time. Along with this expected change, the psychometric properties of the PROM may no more fit to the patients' clinical state and, therefore, the risk of a significant ceiling effect may also be expected to increase after a long follow-up period. This may cause problems in studies comparing outcomes between different patient groups using a PROM, as the gradual concentration of scores toward the extremes of the PROM scale in all patient groups leads to a shrinkage of the observable differences in scores and in the clinical states of the groups (Fig 1). In clinical research, this may lead to concluding that there is no difference in clinical states between two treatment groups when there is, in fact, a difference.

The aim of this methodological review was to assess the influence of the ceiling effect on the conclusions and the risk of an unjustified ''no difference'' conclusion in orthopedic randomized controlled trials (RCTs) reporting no differences in PROM scores between the compared patient groups. We hypothesized that the prevalence of a ceiling effect would be high, leading to a high prevalence of unjustified ''no difference'' conclusions.

2. Methods

2.1. Systematic review

The study was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [6]. A systematic review of RCTs published in the 10 highest-ranking orthopedic journals (2021 Clarivate Analytics) in 2022 was conducted. The following journals were included: *Arthroscopy*, *Bone & Joint Research*, *Journal of Orthopaedics and Traumatology*, *Journal of Orthopaedic Translation*, *Osteoarthritis and Cartilage*, *Spine*, *The American Journal of Sports Medicine, The Bone & Joint Journal*, *The Journal of Arthroplasty*, and *The Journal of Bone and Joint Surgery.*

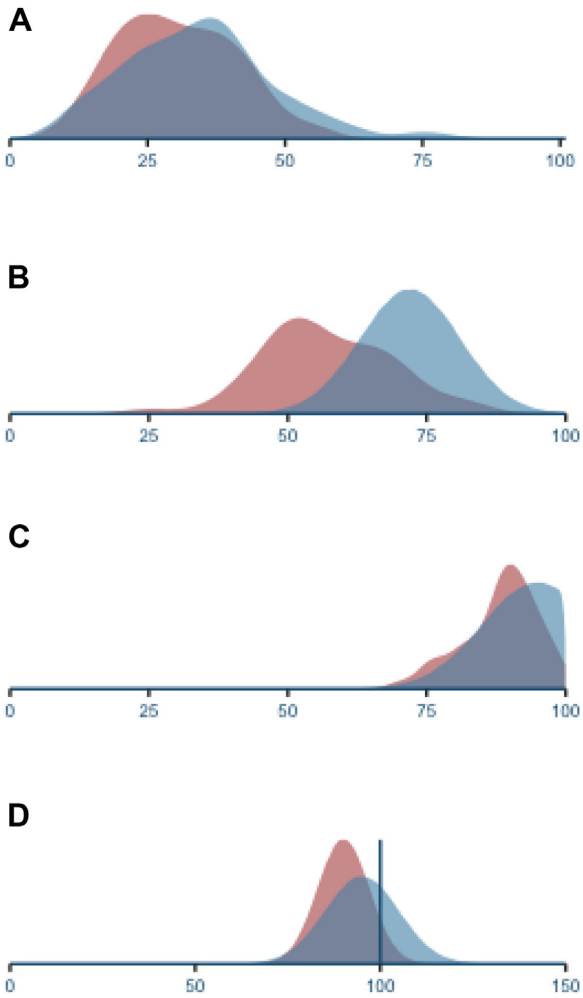

Figure 1. Examples of the presence of floor and ceiling effects. A, Data distributions of two groups that are very similar, with means situated in the middle of the scale and no significant floor or ceiling effects. In this scenario, there is no statistically significant difference between the two groups, and any observed differences are likely due to chance. The distributions in (B) are separated, suggesting a statistically significant difference between the two groups, thus indicating that the groups may differ at the population level and that the observed differences are not due to chance. C, Significant ceiling effect in both groups, with no statistically significant difference between them. D, Simulated normal distributions of unbounded scale with the same means and standard deviations as in (C), and the proportions of the patients exceeding the scale in the simulated dataset differ between the groups. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Eligible articles were identified through PubMed and Scopus (Elsevier) searches using journal filtering. Preliminary screening was performed by two observers independently using titles and abstracts. A third observer was involved in cases of conflict. The inclusion criteria were an RCT study design, a comparison between two parallel study groups, a PROM used as a primary or secondary outcome measure, a parametric statistical test used, and reporting no significant differences in PROM scores between the two study groups as the study outcome. Studies comparing more than two groups were excluded. Studies reporting visual analog scale (VAS) as the outcome were excluded. The search strategy is presented in supplementary materials. Study protocol was registered to Prospero [7].

Full texts were screened by two authors. Articles were included if their statistical methods requiring a normal distribution were used. Articles reporting no differences based on nonparametric tests were excluded. Articles with missing information were excluded.

The data collected from the included studies included the journal names, PROMs used, scales of the PROMs, sample sizes, standard deviations, 95% confidence intervals, standard errors, statistical tests used, and *P* values. If no difference was found using multiple PROMs, the first one was included in the analysis. The scores of the latest follow-up time point were collected.

2.2. Statistical analysis

To quantify the potential risk of erroneous conclusions with the presence of ceiling effect, we conducted an analysis on simulated data based on the dispersion parameters reported in the included studies. Our analysis was based on the following assumptions:

1. The qualities measured by PROMs are continuous and linear by their nature in real life.
2. These qualities approximate to normal distribution in the population from which the study sample is derived.
3. The dispersion parameters (mean, SD, variance, 95% confidence interval) observed in the study sample approximate to those in the population.

Based on these premises, we calculated the proportions of patients that potentially exceed the measurement scale of the PROM. The proportion reflects the proportion of patients for whom the PROM may not be considered a reliable representation of their clinical states. The higher the proportion, the more uncertain is the real-life effect estimation.

The statistical analysis was performed using R 4.2.3 (R Foundation for Statistical Computing, Vienna, Austria). PROM scores were converted to a scale from 0 to 100, with higher values indicating better outcomes. Standard deviations were calculated if only a 95% confidence interval or standard error was reported. By assuming that the traits measured by PROMs are continuous and normally distributed and that these traits can extend beyond the PROM scales, we calculated the proportion potentially surpassing the scale limits using the observed means and standard deviations. Simulated datasets following normal distribution without scale boundaries were generated using the observed sample means and standard deviations [$X \sim N(\mu, \sigma)$, where μ = observed sample mean, σ = observed sample standard deviation; Fig 1C&D]. The proportions of patients scoring above the scale limits (>100 in the simulated datasets) were calculated using the cumulative distribution function

of the normal distribution. The central 95% of the theoretical score distributions were determined as the mean ± 1.96 × SD. When calculated in this manner, the sample mean and standard deviation are likely to be underestimated due to the limited scale of the PROMs in cases of ceiling effect. Consequently, the maximum proportion of patients exceeding the scale is restricted to between 0 and 50%. This restriction arises because, within the PROM score distribution, the highest achievable mean score is equivalent to the maximum score on the PROM scale, implying that only up to 50% of patients can score above the mean value.

To demonstrate the risk of an unjustified ''no difference'' conclusion, we calculated 'a discrepancy coefficient' for each study. The discrepancy coefficient was calculated as follows: the proportions of patients exceeding the scales were compared by calculating a ratio between the study groups using a standard logarithm [$|\log_{10}$ (group1 ÷ group2)|] and the resulting logarithm value was multiplied by the overall proportion of patients exceeding the scale to account the severity of the overall ceiling effect. A higher value of the discrepancy coefficient indicates a greater discrepancy between the proportions of the study groups exceeding the scale and a higher probability of a real-life difference in an outcome, which is suggestive of an unjustified ''no difference'' conclusion despite the nonsignificant result in statistical analysis. With respect to the proposed cutoff for a significant ceiling effect at 15% [8], we established cutoffs for the discrepancy coefficient, indicating potential and increased risk for unjustified conclusions. These cutoffs are determined by the logarithm representing the discrepancy between the proportions of the study groups exceeding the scale by 2:1 and 3:1, respectively, multiplied by 0.15 [log10 (2 ÷ 1) × 0.15 = 0.045 for potential risk and log10 (3 ÷ 1) × 0.15 = 0.072 for increased risk].

3. Results

The initial search identified 2849 studies. After screening, 190 studies involving 24,803 patients were included in the analysis (Fig 2, Supplementary material). The mean sample size was 112 patients (median 85). In most studies (n = 140, 74%), the central 95% theoretical distribution of the scores exceeded the PROM scale (Fig 3). In those studies, the mean proportion exceeding the scale was 22% (median 21%). These proportions exceeded the scales by 10% in 61 studies, by 20% in 33 studies, and by 30% in four studies (Fig 4). Based on the discrepancy coefficients, 39 studies (21%) were identified as having a potential risk, and 15 studies (7.9%) showed an increased risk for unjustifiably concluding 'no difference' (Fig 5).

4. Discussion

The findings of this review suggest that the ceiling effect is a major source of uncertainty in orthopedic RCTs reporting no differences in patient-reported outcomes between the study groups. When comparing treatment options, this uncertainty may lead to fallacies and, consequently, to worse treatment outcomes.

Ceiling effect reduces the power of detecting differences between groups. If a statistical test indicates no significant difference with the presence of a ceiling effect in the groups under comparison, there may indeed be no difference. However, the absence of evidence (supporting a difference) is not evidence of absence [9]. Therefore, in the presence of a ceiling effect, the evidence behind the ''no difference'' conclusion should be critically assessed, as this effect may conceal the difference between the compared groups, thereby obscuring a beneficial treatment effect.

In conclusion, the decision as to whether a predetermined level of subjective well-being is considered satisfactory for patients ultimately becomes a philosophical question. Clinicians often establish objectives for the treatments they administer, aiming to achieve a specific level of functionality or health-related quality of life. Should these objectives be met through the treatment, it can be regarded as successful. Nonetheless, in establishing these predefined objectives, clinicians might unintentionally impose constraints on the potential observable benefits of the

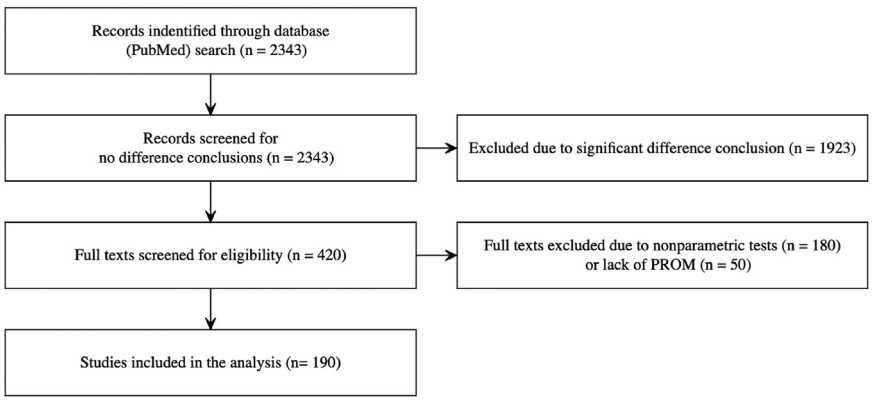

Figure 2. PRISMA flow chart of the review process. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROM, patient-reported outcome measure.

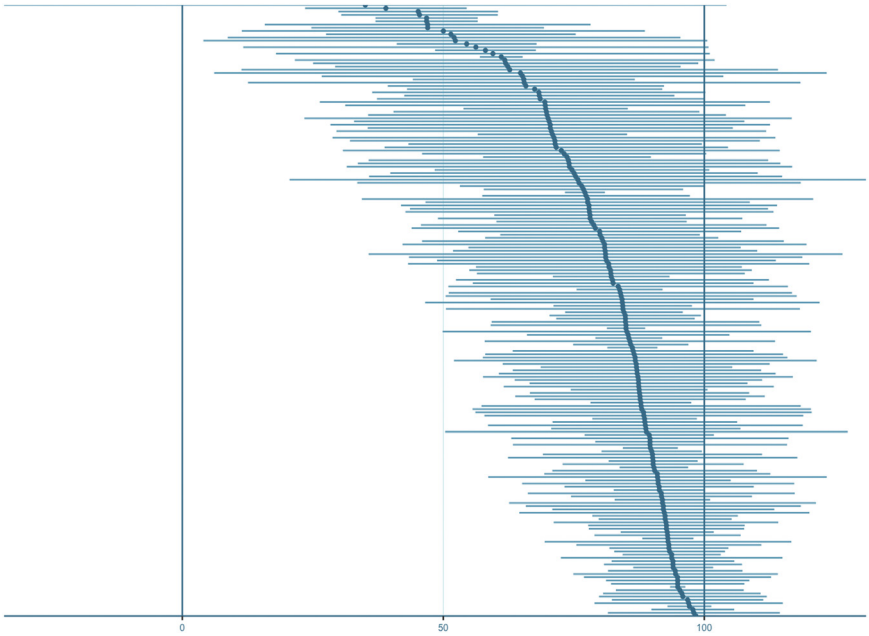

Figure 3. Distributions of the simulated PROM scores based on the observed means and standard deviations. Included studies are on the y-axis and PROM scale from 0 to 100 on the x-axis. The points show the observed means, and the lines show the central 95% theoretical distributions. PROM, patient-reported outcome measure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

treatment. For example, in an acute phase, a patient may not be able to jump and run resulting in a low score in a PROM measuring the ability to jump and run. After the follow-up period, the ability to jump and run may have been restored thereby leading to a maximum score. With regard to jumping and running, the patient may be considered to have completely no disability. However, despite this, the patient may still not be able to play football due to pain or discomfort related to movement. While some may argue that the ability to jump or run should be

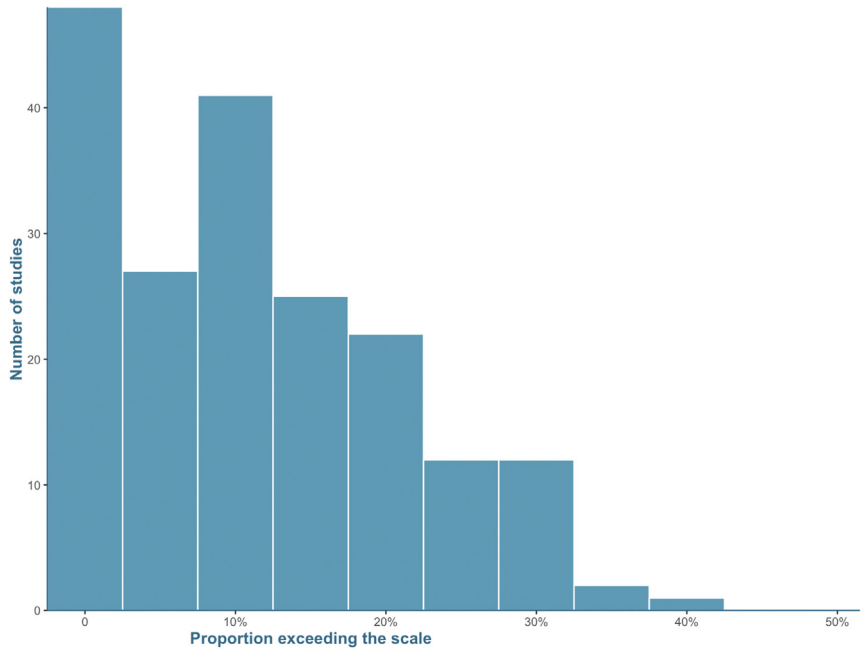

Figure 4. Proportions of patients exceeding the PROM scales. The maximum possible proportion is 50%. PROM, patient-reported outcome measure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

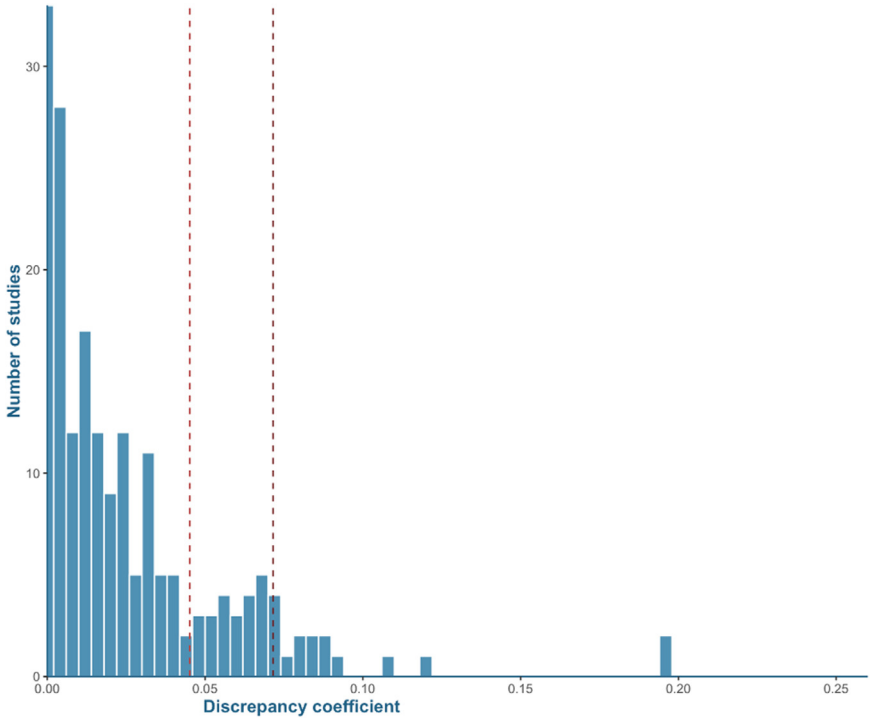

Figure 5. The distribution of discrepancy coefficients per studies. The vertical dashed lines represent the cutoffs for potential (left line) and increased (right line) risk for unjustified ''no difference'' conclusion. In 21% of the studies the risk was moderate and in 7.9% the risk was high. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

considered a satisfactory clinical outcome, and orthopedic treatment may not necessarily target the ability to play football, it is essential to note that when comparing two orthopedic treatments in terms of restoring physical abilities, it would not be justified to conclude that there is no difference between the two treatments if the patients are able to play football in one group and unable to do so in the other. Such a study can determine that both groups of patients achieve a satisfactory level of ability to jump and run, as indicated by the maximum score on the PROM. However, it cannot definitively determine which treatment is superior based on real-life outcomes or whether there are any significant differences between the two groups. Unjustified ''no difference'' conclusions may lead to abandoning of potentially beneficial treatment options, which increases the tendency to favor more conservative solutions.

To address this issue, it is important to choose a suitable and responsive outcome measure that accurately reflects patients' clinical states after an appropriate follow-up period. It is crucial to assess the relevance of using the same outcome measure at multiple time points (eg, 1 month, 6 months, 1 year, and 3 years) following the acute phase, particularly when a significant clinical improvement is expected, regardless of exposure. The presence of a ceiling effect indicates that the study sample exceeds the items' measurement capacity, rendering the PROM unsuitable for evaluating the clinical state of the study sample.

It is the responsibility of researchers to assess the ability of the selected outcome measure to reflect patients' clinical states even at the last follow-up time point. Statistical adjustments have been proposed to reduce the uncertainty caused by the ceiling effect [10]. However, due to the limitations of a PROM beyond its scale range, such adjustments cannot fully resolve the uncertainty arising from the ceiling effect. Therefore, an unexpected ceiling effect should be acknowledged as a potential bias-causing factor, warranting caution in the interpretation of conclusions due to the consequent and inevitable uncertainty. To best avoid these situations, it is recommended to take a meticulous and critical consideration on what would be expected clinical state after the planned follow-up period and select the outcome measure instruments accordingly. Authors should also assess the relevance of the clinical state domains and items of a PROM for patients after the follow-up period. The scarcity of studies evaluating the psychometric properties of PROMs several years after the acute phase highlights a significant knowledge gap in PROM research. There is a pressing need for research focusing on the psychometric properties of PROMs after medium- to long-term follow-up. Additionally, it raises a philosophical and value-based question: whether the maximum score of a PROM truly represents the best imaginable clinical state post-treatment or if advancements in medicine have raised expectations beyond conventional conceptions of desirable treatment outcomes.

The effective presentation of statistical analysis results plays a key role in readers' understanding of a study's topic, and potential flaws in the study methodology often go unrecognized [11,12]. It is important for authors, reviewers, and editors to

actively promote the use of appropriate outcome measures and insist on the accuracy of the derived conclusions. High-impact journals, such as those included in this review, should serve as exemplars of good scientific practices.

4.1. Limitations

We included only RCTs using parametric tests. However, the same uncertainty may be present regardless of the study design or test type. Although we assessed the prevalence of the ceiling effect in orthopedic science, a ceiling effect may be present in any PROM used as the outcome measure. We acknowledge that in some cases, authors may have interpreted the PROM results justifiably, even in the presence of a ceiling effect. However, our intention was to highlight the issue and the associated risk of flawed conclusions due to the ceiling effect. Therefore, we did not analyze the relevance of the final conclusions of the included studies or the relationship between the results obtained with a PROM experiencing a ceiling effect and other supporting evidence. We excluded articles employing nonparametric statistical methods from our analysis to simplify our approach, which assumes a normal distribution of the variable. However, it is important to note that ceiling and floor effects occur irrespective of the validity of the normal distribution assumption. Our statistical approach was based on predefined theoretical assumptions that may not reflect the reality of the patients' clinical states. The rationale was to quantify the extent of potential inferential errors related to the ceiling effect. However, our approach should not be considered as an accurate representation of the qualities measured by the PROMs used in the included studies. We calculated the proportion of patients exceeding the scale using the data obtained from the given PROM. By assessing outcomes using a given PROM, all patients were forced to lie within the scale irrespective of their ''true'' clinical states. Therefore, when simulating the proportions that exceeded the scale, due to the mathematical approach used, the maximum possible proportion exceeding the scale was 50%. However, it is possible that the proportions of patients whose clinical states exceeded the measurement ability of a given PROM were even higher than 50% (up to 100%). Therefore, it is likely that the actual proportion of patients exceeding the scale is higher than the simulated proportion. Thus, the finding of this study may be considered as a bottom line for the underlying risk of errors related to the ceiling effect in PROM scores.

5. Conclusion

It is common to have a mismatch between the chosen PROM and the population being studied increasing the risk of an unjustified ''no difference'' conclusion due to a ceiling effect. The presence of a ceiling effect should be assessed when comparing outcome scores. If a ceiling effect is observed, its presence should be explicitly acknowledged, and it should be clearly stated that no definite conclusions can be drawn.

CRediT authorship contribution statement

Antti Saarinen: Writing – original draft, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Oskari Pakarinen:** Writing – review & editing, Visualization, Validation. **Matias Vaajala:** Writing – review & editing, Visualization, Validation. **Rasmus Liukkonen:** 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, Validation, Supervision, Conceptualization.

Data availability

Data are available for a reasonable request.

Declaration of competing interest

A.S. received financial support from Vappu Uuspää Foundation, and Päivikki and Sakari Sohlberg Foundation. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Heini Huhtala for her critical feedback on the manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111308.

References

[1] MOTION Group. Patient-reported outcomes in orthopaedics. J Bone Joint Surg Am 2018;100:436–42.
[2] Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
[3] Uimonen M, Kuitunen I, Pakarinen O, Vaajala, M., Liukkonen, R., Tukiainen, H., Ponkilainen, V.. Quality of surgical patient-reported outcome measure (PROM) validation studies is often deficient: a systematic review. J Clin Epidemiol.
[4] Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36:648–62.

[5] Wang L, Zhang Z, McArdle JJ, Salthouse TA. Investigating ceiling effects in longitudinal data analysis. Multivariate Behav Res 2008; 43:476−96.

[6] Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264−9.

[7] Prospero ID: CRD42023437721. Available at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=437721.

[8] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293−307.

[9] Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. BMJ 1995;311:485.

[10] Twisk J, Rijmen F. Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. J Clin Epidemiol 2009;62:953−8.

[11] Smith PRM, Ware L, Adams C, Chalmers I. Claims of 'no difference'or 'no effect'in Cochrane and other systematic reviews. BMJ Evid Based Med 2021;26:118−20.

[12] Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open 2019;9:e024785.