ELSEVIER

Journal of Clinical Epidemiology 123 (2020) 27−38

Journal of Clinical Epidemiology

CONTROVERSY & DEBATE SERIES

Controversy and Debate on Meta-epidemiology. Paper 1: Treatment effect sizes vary in randomized trials depending on the type of outcome measure

Dorthe B. Berthelsen[a], Elisabeth Ginnerup-Nielsen[a], Carsten Juhl[b,c], Hans Lund[d], Marius Henriksen[a,e], Asbjørn Hróbjartsson[f], Sabrina M. Nielsen[a], Marieke Voshaar[g], Robin Christensen[a,h,*]

[a]*The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark*

[b]*Research Unit of Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark*

[c]*Department of Physiotherapy and Occupational Therapy, University Hospital of Copenhagen, Herlev and Gentofte, Gentofte, Denmark*

[d]*Centre for Evidence-Based Practice, Western Norway University of Applied Sciences, Bergen, Norway*

[e]*Department of Physical and Occupational Therapy, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark*

[f]*Center for Evidence-Based Medicine, University of Southern Denmark/Odense University Hospital, Odense, Denmark*

[g]*Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands*

[h]*Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Odense, Denmark*

Accepted 23 October 2019; Published online 23 March 2020

Abstract

Objective: To compare estimated treatment effects of physical therapy (PT) between patient-reported outcome measures (PROMs) and outcomes measured in other ways.

Study Design and Setting: We selected randomized trials of PT with both a PROM and a non-PROM included in Cochrane systematic reviews (CSRs). Two reviewers independently extracted data and risk-of-bias assessments. Our primary outcome was the ratio of odds ratios (RORs), used to quantify how effect varies between PROMs and non-PROMs; an ROR $>$ 1 indicates larger effect when assessed by using PROMs. We used REML-methods to estimate associations of trial characteristics with effects and between-trial heterogeneity.

Funding: This study had no specific source of funding. The Parker Institute, Bispebjerg and Frederiksberg Hospital is supported by a core grant from the Oak Foundation (OCAY-13-309); the Oak Foundation had no role in the study design, data collection and analyses, interpretation or reporting of this work, or the decision to submit the work for publication.

Conflict of interest: DBB has nothing to disclose. EG-N has nothing to disclose. CJ has nothing to disclose. HL has nothing to disclose. MH reports personal fees from Thuasne group, outside the submitted work, and is a member of the Association of Danish Physiotherapists that may benefit from the submitted work. AH has nothing to disclose. SMN has nothing to disclose. MV has nothing to disclose. RC reports nonfinancial support from Board membership, grants from Consultancy (AbbVie, Amgen, Axellus A/S, Biogen, Bristol-Myers Squibb, Cambridge Weight Plan, Celgene, Eli Lilly, Hospira, MSD, Norpharma, Novartis, Orkla Health, Pfizer, Roche, Sobi Takeda), personal fees from Employment (Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark), nonfinancial support from Expert testimony, grants from Grants/grants pending (Axellus A/S, AbbVie, Cambridge Weight Plan, Janssen, MSD, Mundipharma, Novartis, and Roche), grants from Payment for lectures including service on speakers bureaus (Abbott, Amgen, Axellus, Bayer HealthCare Pharmaceuticals, Biogen Idec, Bristol-Myers Squibb, Cambridge Weight Plan, Ipsen, Janssen, Laboratoires Expanscience, MSD, Mundipharma, Norpharma, Novartis, Pfizer, Roche, Rottapharm-Madaus, Sobi, and Wyeth), grants from Payment for manuscript preparation (Axellus, Bristol-Myers Squibb, and Cambridge Weight Plan, Aleris-Hamlet (via Norpharma)), nonfinancial support from Patents (planned, pending, or issued), nonfinancial support from Royalties, grants from Payment for development of educational presentations (Bristol-Myers Squibb, MSD, Pfizer), nonfinancial support from Stock/stock options, grants from Travel/accommodations/meeting expenses unrelated to activities listed (Abbott, AbbVie, Axellus, Biogen, Bristol-Myers Squibb, Cambridge Weight Plan, Celgene, Laboratoires Expanscience, Norpharma, Novartis, Pfizer, Roche, Rottapharm-Madaus, and Wyeth), nonfinancial support from Other (err on the side of full disclosure), outside the submitted work; RC is involved in many health-care initiatives and research that could benefit from wide uptake of this publication (including Cochrane, OMERACT, IDEOM, RADS, and the GRADE Working Group). Musculoskeletal Statistics Unit, The Parker Institute is grateful for the financial support received from public and private foundations, companies and private individuals over the years. The Parker Institute is supported by a core grant from the Oak Foundation; The Oak Foundation is a group of philanthropic organizations that, since its establishment in 1983, has given grants to not-for-profit organizations around the world.

* Corresponding author. Biostatistician & Professor, Head of Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksberg Hospital, Nordre Fasanvej 57, DK-2000 CopenhagenF, Denmark. Tel.: (+45) 3816 4165; fax: (+45) 3816 4159.

E-mail address: Robin.Christensen@regionh.dk (R. Christensen).

https://doi.org/10.1016/j.jclinepi.2019.10.016

Results: From 90 relevant CSRs, 205 PT trials were included. The summary ROR across all the comparisons was not statistically significant (ROR, 0.88 [95% CI: 0.70−1.12]; P = 0.30); however, the heterogeneity was substantial (I^2 = 88.1%). When stratifying non-PROMs further into clearly objective non-PROMs (e.g., biomarkers) and other non-PROMs (e.g., aerobic capacity), the PROMs appeared more favorable than did clearly objective non-PROMs (ROR, 1.92 [95% CI: 0.99−3.72]; P = 0.05).

Conclusion: Estimated treatment effects based on PROMs are generally comparable with treatment effects measured in other ways. However, in our study, PROMs indicate a more favorable treatment effect compared with treatment effects based on clearly objective outcomes.

Keywords: Patient-reported outcome measures; Effect size; Patient-involvement; Physical therapy; Meta-epidemiology; Meta-Research

1. Introduction

Patients' self-reports can greatly help their clinician treat them. Such patient-reported outcome measures (PROMs) tell how patients function or feel in relation to a health condition and its therapy, without interpretation by the clinician or anyone else [1]. But who provides more reliable information for basing treatment: the patient or the highly trained clinician? What sort of information is most useful for producing favorable treatment outcomes?

PROMs are frequently referred to as subjective [2], potentially leading to an increased risk of ascertainment bias when patients are their own outcome assessors [3,4], and to performance bias if blinding is not possible [5,6]. Thus, when blinding is impossible in randomized controlled trials (RCTs), outcomes not susceptible to patient behavior are recommended [5]. Both patients' and clinicians' subjectively assessed treatment effects are more prone to bias, leading to larger effect sizes, than are ‘‘objectively assessed'' effects of treatment [7−9]. However, patients' self-reports (i.e., PROMs) of effects are often different from clinicians' assessments, even when based on the same outcome measure [10−12]. Still, Evangelou et al. found no difference between doctors' and patients' global assessments of treatment effects [13]. Nevertheless, Cohen et al. indicated that PROMs demonstrated better discrimination of effects than did clinician-reported outcomes in rheumatoid arthritis patients treated with anti-IL1 [14].

Although objective non-PROMs might only represent biomarkers or surrogate outcomes [15,16], they are less susceptible to bias when the assessor is blinded. Conversely, PROMs can increase clinical relevance—but with an increased risk of performance bias. Patients, ultimately, would benefit if clinicians and researchers knew whether PROMs or objective non-PROMs led to more favorable clinical effects (e.g., when developing core outcome measurement sets in various conditions [17,18]).

The aim of this meta-epidemiological study was to explore whether estimates of treatment effect differ between PROMs and non-PROMs in RCTs evaluating nonpharmacological interventions that can be difficult to blind, using PT as such an example [19]. We hypothesized that treatment effects of PT would show more favorable estimates when assessed by PROMs than by non-PROMs.

2. Methods

The protocol of this study (details in Appendix 1) is registered on PROSPERO (CRD42017055974). This study is reported according to the PRISMA statement [20].

2.1. Patient involvement

To ensure that the study objectives was assessed from the patient's point of view, three patient research partners (PRPs) were involved in designing the study and discussing the results. We identified the PRPs through the Danish Multiple Sclerosis Society, The Danish Rheumatism Association, and the Danish Heart Foundation. Their involvement was based on the European League Against Rheumatism recommendations for including patient representatives in scientific projects [21].

2.2. Data sources and searches

We searched the Cochrane Database of Systematic Reviews via the Cochrane Library. Our search strategy, previously described by Ginnerup−Nielsen et al., represented multiple definitions of search terms related to PT [22] (Appendix 1). We had no limit for publication year. We conducted our search on February 22, 2017.

2.3. Study selection

Eligible trials had to be included in a Cochrane systematic review (CSR) evaluating a PT intervention, and the CSR had to include separate meta-analyses (presented as forest plots) of PROMs and non-PROMs for comparing the same PT intervention against the same control group.

To identify relevant CSRs, we excluded protocols and withdrawn reviews. Two reviewers (DBB/EG-N) independently assessed the retrieved CSRs for eligibility by screening title and abstract and for inclusion of figures, as reviews without figures would not include a forest plot. Disagreements were resolved by discussion between the two reviewers or by involving a third reviewer (RC). To capture

What is new?

Key findings

- Patients' self-reports of treatment effects from physical therapy (PT) are generally comparable with outcomes measured in other ways.
- Treatment effects appear more favorable when assessed by PROMs than by clearly objective outcomes such as biomarkers.
- When outcomes reflect the same construct, PROMs appear less favorable than comparable non-PROMs.

What this adds to what was known?

- Patients and clinicians' different perspectives on a disease may influence estimates of treatment effects in randomized trials.

What is the implication and what should change now?

- Researchers should consider using objective outcome measurements together with PROMs to cover the pathophysiological manifestations; and this should also be taken into account when developing Core Outcome Sets for various conditions.

eligible meta-analyses, we examined figures and tables of relevant CSRs, and if in doubt, we examined the full text. We accepted experimental interventions compared with placebo, sham, usual care, or no intervention. We accepted meta-analyses regardless of the number of trials included or whether the outcomes were binary or continuous. From each CSR, we selected one meta-analysis reporting a PROM and one meta-analysis reporting a non-PROM. If there were multiple obvious comparisons of meta-analyses, we selected the meta-analyses including the largest numbers of trials; if an equal number of trials was included, we selected the first reported meta-analysis of eligible comparisons. We chose comparisons for the same time point. Individual trials were selected if estimates from the same trial were included in the meta-analysis of both the PROM and the non-PROM.

2.4. Data extraction and quality assessment

Two reviewers (DBB/EG-N) used a standardized form to independently extract data and retrieve risk of bias (RoB) from the eligible CSRs. Disagreements were resolved through discussion between the two reviewers or by the involvement of a third reviewer (RC).

For each trial, we extracted information about the author, year of publication, population, intervention, and control. For each outcome, we extracted estimates, confidence intervals, scale of measurement, and number of patients in the intervention and the control groups.

We collected data on trials from the included meta-analyses, but when a description was unclear, we obtained data from published articles. Domains of selection, performance, and attrition bias for each trial were retrieved from the CSR author's judgment (Appendix 2) using the Cochrane Collaboration's RoB tool [23,24].

2.5. Data synthesis and analysis

For each extracted effect estimate (i.e., contrast for experimental intervention vs. control comparator) in each trial, we estimated effects as odds ratios (ORs). Continuous outcomes were converted via the standardized mean difference (SMD) to $\log_e$ORs by multiplying the SMD by $\pi/\sqrt{3}$, as described by Chinn [25]. Outcomes were recoded so that an OR >1 indicated a more favorable effect of the experimental intervention relative to the control comparator. Subsequently, within each trial, we estimated the differences in estimates ($\log_e[\mathrm{OR}_{\mathrm{PROM}}]$ vs. $\log_e[\mathrm{OR}_{\text{non-PROM}}]$) by using the meta-epidemiological approach described by Sterne et al. [26], calculating the ratio of ORs (ROR) for each comparison ($\mathrm{ROR} = \mathrm{OR}_{\mathrm{PROM}}/\mathrm{OR}_{\text{non-PROM}}$). Thus, an estimate of ROR >1 indicates a more favorable effect assessed by PROMs over non-PROMs. The correlation between the two $\log_e$ORs was estimated empirically across all comparisons and subsequently applied to estimate the corresponding variance of ROR, as the OR measures were not mutually independent (i.e., the paired OR measures came from the same trial). To combine the individual RORs across trials, we used mixed-effects (restricted maximum likelihood) meta-regression methods to estimate the between-study variance and the combined estimate [27]. We intended to adjust for the individual CSRs. However, because of collinearity, this was not feasible. We used the Cochran's Q test and the I^2 metric to quantify heterogeneity and inconsistency across the estimated RORs [28,29].

2.6. Stratified and sensitivity analyses

We categorized population in broad groups according to the International Statistical Classification of Diseases [30]. Interventions were identified and categorized according to the World Confederation for Physical Therapy [31] and the American Physical Therapy Association [32]. We also classified interventions to be either active treatments (i.e., intervention mainly based on the patient's being active [e.g., exercising]) or passive treatments (i.e., interventions mainly done by the physiotherapist to the patient [e.g., manual therapy]). Control comparator groups were categorized as placebo/sham or usual care/no intervention.

For our primary analysis, we included all eligible comparisons. We carried out sensitivity analyses using

DerSimonian and Laird random-effects, and fixed-effects meta-analyses.

To assess the influence of study characteristics on treatment effects, we undertook prespecified stratified analyses according to classification of disease, intervention (including active or passive treatments), control group, and whether estimates differed in subgroups of trials according to overall RoB and each domain of RoB.

Additional stratified analyses were performed according to whether the trial was included in a CSR published before or after 2009 (the Cochrane RoB tool was introduced in 2008 [23]). Post hoc analyses took into account numerous considerations: the scale of measurement used (depending on whether both PROMs and non-PROMs were binary, continuous, or a combination of binary and continuous outcomes); whether PROMs and non-PROMs were measuring clearly the same construct (e.g., mouth dryness vs. unstimulated whole saliva), or measuring not clearly the same construct (i.e., measuring clearly not the same construct [e.g., pain vs. range of motion] or whether it was unclear if the same construct was measured [e.g., Dizziness Handicap Inventory vs. Dynamic Gait Index]); the categories of outcomes merged into 8 overall categories for the PROMs (i.e., quality of life, pain, symptom score, fatigue, physical function, depression and anxiety, dyspnea and other), and 13 categories for the non-PROMs (i.e., aerobic capacity/physical fitness, strength, mortality, clinician-assessed scores and symptoms, range of motion, lung function, micturition/leakage/incontinence, cesarean section, biomarkers, use of analgesics, length of hospital stay, days of work/sick leave, and other); and to classification of clearly objective and less objective non-PROMs. We considered mortality and biomarkers to be *clearly objective* non-PROMs. Less objective non-PROMs included *not objective* non-PROMs (i.e., clinician-assessed scores and symptoms, use of analgesics, length of hospital stay, and days of work/sick leave), and *unclear* non-PROMs (i.e., aerobic capacity/physical fitness, strength, range of motion, lung function, micturition/leakage/incontinence, cesarean section, and other).

Among all our analyses, we considered covariates to be potentially relevant if they demonstrated a statistically significant ability to decrease the between-study variability [tau^2] across strata, and we restricted presentation of subgroup analyses to these covariates. We evaluated the influence of small sample size on estimated ROR by funnel plot inspection [33]. All analyses were performed in STATA version 14.2.

3. Results

3.1. Eligible reviews and trials

As illustrated in Fig. 1, our search identified 456 publications (Appendix 3), of which 417 were deemed potentially eligible after removing protocols and withdrawn reviews. After screening titles and abstracts and checking for inclusion of figures and subsequent forest plots and adherence to our eligibility criteria, we narrowed the field to 101 eligible reviews. Of the eligible reviews, 90—comprising 209 trials—contained meta-analyses of both PROMs and non-PROMs. Four trials were not included in the final analysis, as their effect sizes were deemed "not estimable" in the meta-analyses, yielding a total of 205 trials that were included in the final analysis. There was a high level of agreement (89% [405/456], kappa = 0.71) between the two reviewers in selecting eligible reviews.

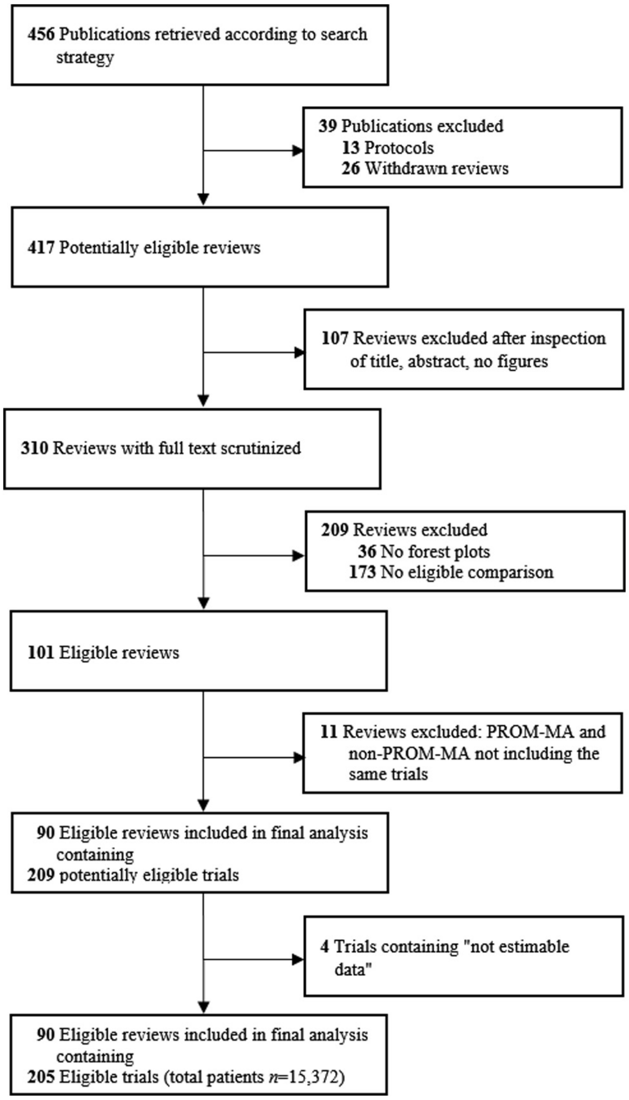
456 Publications retrieved according to search strategy
39 Publications excluded
13 Protocols
26 Withdrawn reviews
417 Potentially eligible reviews
107 Reviews excluded after inspection of title, abstract, no figures
310 Reviews with full text scrutinized
209 Reviews excluded
36 No forest plots
173 No eligible comparison
101 Eligible reviews
11 Reviews excluded: PROM-MA and non-PROM-MA not including the same trials
90 Eligible reviews included in final analysis containing
209 potentially eligible trials
4 Trials containing "not estimable data"
90 Eligible reviews included in final analysis containing
205 Eligible trials (total patients n=15,372)

Fig. 1. Flow diagram for the study selection. *Abbreviations:* PROM-MA, meta-analysis using patient-reported outcome measure; non-PROM-MA, meta-analysis using non-patient-reported outcome measure.

3.2. Characteristics of included trials

Table 1 shows characteristics of the 205 included trials (see also Appendix 4). Patients suffering from conditions

Table 1. Characteristics of included trials

Characteristic	No. of trials ($k = 205$)[a]
Classification of disease	
Musculoskeletal system	40 (19.5)
Respiratory system	36 (17.6)
Genitourinary system/pregnancy	33 (16.1)
Neoplasms	24 (11.7)
Nervous system	22 (10.7)
Circulatory system	20 (9.8)
Conditions related to external causes/ injury	7 (3.4)
Mental disorders	5 (2.4)
Metabolic diseases	1 (0.5)
Other	17 (8.3)
Interventions	
Exercise	111 (54.1)
Manual therapy	24 (11.7)
Physical agents/mechanical modalities	22 (10.7)
Electrotherapeutic modalities	13 (6.3)
Education	7 (3.4)
Devices and equipment	5 (2.4)
Functional training in self-care	3 (1.5)
Airway clearance	2 (1.0)
Functional training in work	1 (0.5)
Integumentary repair/protection techniques	0 (0.0)
Psychomotor therapy	0 (0.0)
Combination	17 (8.3)
Type of intervention	
Active	134 (65.4)
Passive	71 (34.6)
Controls	
Usual care/no intervention	169 (82.4)
Placebo/sham	36 (17.6)
Sample size	
Total no. of patients in intervention groups (non-PROM)	7,841 (51.0)
Total no. of patients in control groups (non-PROM)	7,531 (49.0)
No. of patients in intervention groups, median (IQR)	26 (13–42)
No. of patients in control group, median (IQR)	24 (14–40)
Scale of measurement	
No. of selected binary outcomes	84 (20.5)
No. of selected continuous outcomes	326 (79.5)
Categories of PROMs	
Quality of life	70 (34.1)
Pain	56 (27.3)
Symptom score	28 (13.7)
Fatigue	20 (9.8)
Physical function[b]	13 (6.3)
Depression and anxiety[c]	10 (4.9)
Dyspnea	7 (3.4)
Other	1 (0.5)
Categories of non-PROMs	
Aerobic capacity/physical fitness	66 (32.2)
Strength	21 (10.2)
Mortality	18 (8.8)
Clinician-assessed scores and symptoms	17 (8.3)
Range of motion	16 (7.8)
Lung function	15 (7.3)
Micturition/leakage/incontinence[d]	12 (5.9)
Cesarean section	7 (3.4)
Biomarkers	6 (2.9)
Use of analgesics[e]	6 (2.9)
Length of hospital stay	4 (2.0)
Days of work/sick leave	3 (1.5)
Other	14 (6.8)

Abbreviation: IQR, interquartile range.

[a] Data are expressed as number (%) unless otherwise indicated.

[b] Based on patients' reports such as physical function domains of Short Form 36/Short Form 12.

[c] Based on self-rating scales.

[d] Based on quantifications of symptoms (e.g., number of leakage episodes in 24 hours).

[e] Based on quantitative assessment (e.g., capsule count).

related to the musculoskeletal system ($k = 40$, 19.5%) were the most frequent population studied, whereas therapeutic exercise was the most common intervention studied ($k = 111$, 54.1%). Most trials used usual care/no intervention rather than placebo/sham as control (82.4% vs. 17.6%), and most interventions were active rather than passive (65.4% vs. 34.6%). The median sample size of the included trials was 26 (interquartile range [IQR], 13–42) across the intervention groups and almost identical with the control groups (24 [IQR, 14–40]). Continuous scales were used in 172 (83.9%) of the reported PROMs and 154 (75.1%) of the reported non-PROMs. The most frequently reported PROM was measuring quality of life (34.1%), whereas measuring aerobic capacity (32.2%) was the most frequently reported non-PROM. Trials were published between 1979 and 2016.

Table 2 shows the RoB in the included trials (see also Appendix 4). More PROMs than non-PROMs were considered to present a high risk for performance bias ($k = 121$, 59.0% vs. $k = 33$, 16.1%, $P < 0.001$), whereas the rate of high risk for attrition bias was comparable between treatment effects reported on PROMs and non-PROMs ($k = 28$, 13.7% vs. $k = 30$, 14.6%, $P = 0.78$). Consequently, according to the overall RoB assessment, more PROMs were considered high RoB compared with non-PROMs ($k = 132$, 64.4% vs. $k = 61$, 29.8%, $P < 0.001$).

Table 2. Risk of bias in the included trials[a]

Characteristic	No. of trials including PROMs (k = 205)[a]	No. of trials including non-PROMs (k = 205)[a]
Risk of bias		
Selection bias		
Low	68 (33.2)	68 (33.2)
Unclear	130 (63.4)	130 (63.4)
High	7 (3.4)	7 (3.4)
Performance bias		
Low	19 (9.3)	104 (50.7)
Unclear	65 (31.7)	68 (33.2)
High	121 (59.0)	33 (16.1)
Attrition bias		
Low	130 (63.4)	127 (62.0)
Unclear	47 (22.9)	48 (23.4)
High	28 (13.7)	30 (14.6)
Overall risk of bias		
Low	9 (4.4)	33 (16.1)
Unclear	64 (31.2)	111 (54.1)
High	132 (64.4)	61 (29.8)

Abbreviations: PROM, patient-reported outcome measure; non-PROM, non-patient-reported outcome measure.

[a] Data are expressed as number (%).

3.3. Differences in treatment effects between PROMs and non-PROMs

As shown in Fig. 2, there was no statistically significant summary ROR across the 205 comparisons (ROR, 0.88 [95% CI, 0.70 to 1.12]; P = 0.30), and the heterogeneity among the included trials was high (I^2 = 88.1%; P < 0.001). In the 24 trials using clearly objective non-PROMs, treatment effects appeared more favorable when reported on PROMs compared with clearly objective non-PROMs (ROR, 1.92 [95% CI, 0.99 to 3.72]; P = 0.05). In the 181 trials reporting other less objective non-PROMs, treatment effects reported on PROMs appeared less favorable (ROR, 0.80 [95% CI, 0.62 to 1.02]; P = 0.07). Table 3 (and Appendix 5) shows that in the 24 trials where comparisons were assessing clearly the same construct, treatment effects appeared less favorable using PROMs compared with non-PROMs (ROR, 0.29 [95% CI, 0.15 to 0.55]; P < 0.001). No statistically significant differences were found in the 181 trials where comparisons were assessing not clearly the same construct (ROR, 1.03 [95% CI, 0.81 to 1.31]; P = 0.80).

3.4. Sensitivity analyses

Table 3 shows the results of the sensitivity analyses. The results of the DerSimonian and Laird random-effects, and the fixed-effects meta-analyses were consistent with our primary analysis. Meta-regression examining the influence of intervention, interventions being either active or passive, and controls being either placebo/sham or usual care/no intervention, showed no statistically significant interaction (i.e., the between-study variance was not reduced). However, the interaction among trials of different classifications of diseases and estimates was statistically significant (P = 0.001).

Our post hoc analyses showed that type of non-PROM had a statistically significant interaction with the estimates (P = 0.03). The subsequent dichotomization into clearly objective or not-clearly objective non-PROMs showed similar results (P = 0.02). Furthermore, we observed a statistically significant interaction between estimates and whether PROMs and non-PROMs were measuring clearly the same construct (P < 0.001). We did not observe any statistically significant association with estimates and category of PROM, year of review publication, or scale of measurement.

Subgroup analysis of trials in which the non-PROMs were clearly objective (ROR, 1.70 [95% CI, 1.21 to 2.38]; P = 0.003; I^2 = 48.1; P = 0.01; Appendix 6) showed results similar to those of our meta-regression analyses (Table 3). Conversely, results of subgroup analysis of trials in which comparisons were assessing clearly the same construct (ROR, 0.28 [95% CI, 0.06 to 1.38]; P = 0.11; I^2 = 97.9; P < 0.001; Appendix 6) showed no statistically significant effect, which was inconsistent with the findings of our meta-regression analyses (Table 3). Funnel plots including all trials, and funnel plots including each classification of disease separately were symmetrical on visual inspection, suggesting no presence of small-study effects in the analyses (Appendix 7).

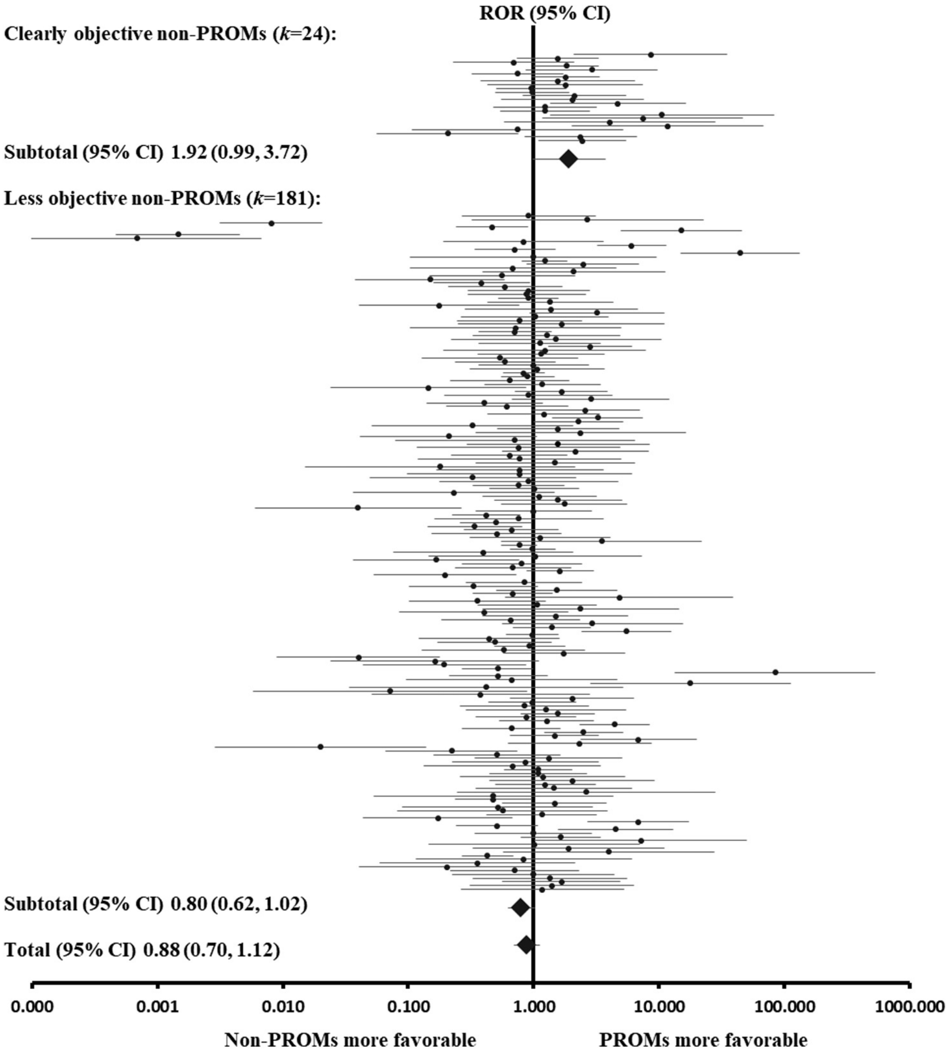

Fig. 2. Results of stratified analysis according to subgroups of trials including clearly objective non-PROMs and less objective non-PROMs. *Abbreviations:* RORs, ratio of odds ratios; PROMs, patient-reported outcome measures.

3.5. Risk of bias

As represented in Table 3, the sensitivity analyses examining the influence of RoB in the included RCTs showed no significant reduction of the between-study variance between estimates and trial quality. There was no statistically significant difference between estimates assessed by PROMs or non-PROMs in trials of low RoB (ROR, 1.12 [95% CI, 0.32 to 3.92]; $P = 0.85$) or trials of high RoB (ROR, 0.86 [95% CI, 0.64 to 1.14]; $P = 0.29$).

Post hoc analysis of RoB in subgroups of trials reporting clearly objective non-PROMs and less objective non-PROMs, respectively, was not robust enough to explain the differences in RoB (Appendix 8).

4. Discussion

In this meta-epidemiological study, we compared treatment effects of PT according to reporting of PROMs and non-PROMs in RCTs. We used a sample of 205 trials from 90 CSRs representing a wide range of populations and PT interventions. We found no overall difference between treatment effects assessed by PROMs and non-PROMs. As prespecified, we analyzed whether our result was influenced by population, intervention (including passive or active treatment), control group, or RoB. Different populations of diseases only partially explained our findings. We explored the influence of publication year, scale of measurement, comparisons' assessment of same constructs, and overall categories of PROMs and non-PROMs, including subgroups of non-PROMs reflecting clearly objective non-PROMs (i.e., mortality and biomarkers) and less objective non-PROMs. When comparisons assessed the same construct, treatment effects appeared less favorable using PROMs rather than non-PROMs. However, when the non-PROM was clearly objective, treatment effects reported on PROMs appeared more favorable. On the contrary, patients' own report of treatment effects

Table 3. Results of the stratified analyses

Comparisons	No. of trials	ROR	95% CI	I^2	Tau^2	*P* For interaction
Primary analysis (REML)	205	0.88	0.70, 1.12	88.1	2.385	-
Random-effects meta-analysis[a]	205	0.89	0.72, 1.09		(1.882)	-
Fixed-effects meta-analysis	205	0.93	0.87, 0.99			-
Classification of disease				86.1	2.130	0.001
Musculoskeletal system	40	0.98	0.59, 1.62			
Respiratory system	36	1.11	0.65, 1.90			
Genitourinary system/pregnancy	33	1.37	0.80, 2.33			
Neoplasms	24	0.69	0.36, 1.32			
Nervous system	22	0.84	0.43, 1.67			
Circulatory system	20	0.62	0.31, 1.25			
Conditions related to ext. causes/injury	7	0.67	0.18, 2.43			
Mental disorders	5	0.04	0.01, 0.15			
Metabolic diseases	1	0.18	0.00, 8.26			
Other	17	1.26	0.56, 2.84			
Intervention				88.3	2.401	0.450
Exercise	111	0.78	0.57, 1.06			
Manual therapy	24	0.59	0.30, 1.17			
Physical agents/mechanical modalities	22	1.35	0.65, 2.80			
Electrotherapeutic modalities	13	1.37	0.55, 3.43			
Education	7	1.45	0.43, 4.85			
Devices and equipment	5	0.89	0.21, 3.76			
Functional training in self-care	3	2.03	0.31, 13.26			
Airway clearance	2	8.06	0.70, 92.62			
Functional training in work	1	0.44	0.02, 12.20			
Combination	17	0.87	0.39, 1.91			
Type of intervention				88.2	2.391	0.299
Active	71	0.81	0.61, 1.08			
Passive	134	1.05	0.71, 1.56			
Controls				88.1	2.394	0.612
Usual care/no intervention	167	0.86	0.67, 1.11			
Placebo/sham	36	1.01	0.58, 1.78			
Overall risk of bias				88.2	2.411	0.901
Low	7	1.12	0.32, 3.92			
Unclear	63	0.92	0.61, 1.40			
High	135	0.86	0.64, 1.14			
Selection bias				88.2	2.406	0.796
Adequate	68	0.81	0.54, 1.20			
Unclear	130	0.94	0.70, 1.26			
Inadequate	7	0.72	0.20, 2.62			
Blinding, patients and personnel				88.2	2.410	0.884
Adequate	19	0.77	0.35, 1.70			
Unclear	65	0.95	0.63, 1.43			
Inadequate	121	0.87	0.64, 1.18			
Blinding, outcome assessor				88.1	2.393	0.503
Adequate	104	0.99	0.72, 1.37			
Unclear	68	0.85	0.56, 1.27			
Inadequate	33	0.67	0.37, 1.20			
Attrition bias, PROM				88.1	2.370	0.176
Adequate	130	0.85	0.63, 1.13			

(*Continued*)

Table 3. Continued

Comparisons	No. of trials	ROR	95% CI	I^2	Tau^2	*P* For interaction
Unclear	47	1.25	0.77, 2.02			
Inadequate	28	0.60	0.32, 1.14			
Attrition bias, non-PROM				88.1	2.377	0.249
Adequate	127	0.84	0.63, 1.13			
Unclear	48	1.22	0.76, 1.98			
Inadequate	30	0.65	0.35, 1.20			
Category of PROM				88.1	2.378	0.428
Quality of life	70	1.12	0.75, 1.66			
Pain	56	0.99	0.64, 1.55			
Symptom score	28	0.50	0.27, 0.93			
Fatigue	20	0.71	0.34, 1.48			
Physical function	13	0.54	0.21, 1.35			
Depression and anxiety	10	1.29	0.46, 3.62			
Dyspnea	7	1.14	0.29, 4.39			
Other	1	0.66	0.03, 17.73			
Category of non-PROM				87.5	2.244	0.030
Aerobic capacity/physical fitness	66	0.55	0.37, 0.82			
Strength	21	0.76	0.37, 1.55			
Mortality	18	1.75	0.84, 3.65			
Clinician assessed scores/symptoms	17	0.92	0.41, 2.03			
Range of motion	16	1.00	0.44, 2.27			
Lung function	15	2.13	0.90, 5.08			
Micturition/leakage/incontinence	12	1.02	0.42, 2.47			
Cesarean section	7	2.08	0.64, 6.71			
Biomarkers	6	2.68	0.65, 11.02			
Use of analgesics	6	0.79	0.21, 2.91			
Length of hospital stay	4	2.20	0.47, 10.35			
Days of work/sick leave	3	1.16	0.20, 6.89			
Other	14	0.38	0.16, 0.88			
Clearly objective vs. less objective non-PROM				87.9	2.327	0.015
Clearly objective non-PROM	24	1.92	0.99, 3.72			
Less objective non-PROM	181	0.80	0.62, 1.02			
Assessment of construct				87.4	2.224	<0.001
Clearly the same construct	24	0.29	0.15, 0.55			
Not clearly the same construct	181	1.03	0.81, 1.31			
Year of review publication				88.1	2.381	0.241
Year <2009	18	1.39	0.64, 3.05			
Year ≥2009	187	0.85	0.66, 1.08			
Scale of measurement				88.2	2.411	0.867
PROMs and non-PROMs binary	17	1.05	0.48, 2.32			
PROMs and non-PROMs continuous	138	0.85	0.64, 1.13			
Combination	50	0.92	0.58, 1.46			

Abbreviations: RORs, ratio of odds ratios; REML, restricted maximum likelihood.

[a] Based on DerSimonian and Laird.

appeared less favorable when compared with less objective non-PROMs.

The covariates included in the stratified analyses only slightly reduced the heterogeneity across our included trials. The uncertainty in the result was large, and for the remaining heterogeneity, we cannot exclude the possibility that unexplained differences between trials may have averaged out on the large sample, leading to no statistically significant influence on the estimate. Nevertheless, our results of RORs not equal to one can be interpreted as true effect difference, as bias, or as a conceptual variation between measures (i.e., different constructs). However, as bias in research is defined as systematic errors that are introduced into sampling or testing by selecting or encouraging one outcome or answer over others, and with no gold standard against which PROM and non-PROM-based estimates of effects can be compared, we are not entirely able−from a philosophy of science perspective - to determine if a deviation between these estimates can be referred to as bias. Furthermore, differences or lack of differences in RORs may have been due to differences in variability of outcomes, as many of the ORs used as effect estimates were calculated from continuous outcomes. Such SMD effects depend both on the estimate of effect and the variability of the outcome.

Our overall result suggesting that treatment effects did not vary with type of outcome was unexpected, as Wood et al. found subjective outcomes associated with larger treatment effects and increased RoB compared with objective outcomes and mortality [7]. Thus, our result does not suggest that PROMs can be used without RoB. However, our main analysis included clearly objective non-PROMs (e.g., mortality) along with other non-PROMs (e.g., clinician-assessed scores and symptoms), which may be considered equally subjective, as these outcome measures involved personal judgment. Our post hoc analysis, indicating more favorable effects of treatment reported on PROMs when the non-PROM was clearly objective, suggest that lack of blinding could have influenced our results. However, we found no statistically significant difference in treatment effects between trials where patients were adequately or inadequately blinded, nor where outcome assessors were adequately or inadequately blinded. Thus, it is unlikely that bias due to lack of blinding may explain our results. Nevertheless, Wood et al. categorized clinician-assessed outcomes as subjective outcomes, whereas we categorized clinician-assessed outcomes as less objective outcomes, which may explain the differences in the results of our study. However, the objective of PT may not always be measurable by a truly objective non-PROM such as mortality or biomarker, which may also explain the observed difference, as—for example—mortality and a PROM most likely do not measure the same construct.

Other studies have examined treatment effects assessed by patients and clinicians; Khanna et al. compared patients' and clinicians' evaluations of the same outcome and found greater pain and worse function when patients assessed effects following total knee arthroplasty, compared with the clinicians' assessed effects [10]. Similar results were found in chronic myeloid leukemia [11] and lupus disease [12]. Our analysis based on 24 trials in which comparisons were assessing clearly the same construct was in accordance with these findings (i.e., treatment effects were more favorable using non-PROMs). However, these results were not statistically significant in our subgroup analysis. Nevertheless, Evangelou et al. found patients' and clinicians' assessments of treatment effects similar in diverse conditions [13]. However, different perspectives of effect may have to be taken into consideration; Yen et al. found that patients based their (self-assessed) scores on psychological and physical well-being, whereas clinicians based their assessments on clinical and physical signs and symptoms [12]. Likewise, Basch et al. found that clinicians' assessments better predict unfavorable clinical events such as adverse symptoms, whereas patients' reports better reflect daily health status [34]. Furthermore, Mukesh et al. compared PROMs with clinician-assessed outcomes of physical examination and photographic assessment in patients with breast cancer and found a weak level of concordance [35]. Such apparent variation in patients' and clinicians' focus on different aspects of a disease may have influenced our results; our analysis based on comparisons assessing clearly the same construct may include different subconstructs reflecting patients' and clinicians' different perspectives, which we have not been able to identify.

An apparent limitation with our primary analysis, comparing PROMs and non-PROMs, is that it could potentially ‘‘metaconfound’’ the reported empirical evidence. That is, we cannot exclude the possibility that the differences in effect could be because PT interventions are generally likely to have larger effects on disability (often measured with PROMs) than on survival (a non-PROM); that is, the association between outcome measure and effect could be confounded by construct. Realizing that it is not possible to determine whether PROMs and non-PROMs are ever really measuring the same construct, it might be argued that it is not possible to know if differences in effect estimates arise because of bias or because they are measuring different things. However, we argue that even with this potential metaconfounding caveat, our work gives support to the fact that Core Outcome Set developers should attempt to comprise at least three of the four core ‘‘areas’’ suggested by OMERACT: death, life impact (all aspects of how a patient feels or functions), and pathophysiologic manifestations (disease-specific clinical and psychological signs, biomarkers, and potential surrogate outcome measures necessary to assess specific effects) [18]. Likewise, in daily practice, both PROMs and non-PROMs should be considered to ensure both patients' and clinicians' perspectives on treatment effect are met.

The PRPs shared this perspective, but they emphasized that non-PROMs were far from always patient-relevant and that clinicians tend to have greater trust in ‘‘objective outcomes’’ than in patients' assessments. This might explain why we found patients own report of treatment effects less favorable compared with less objective non-PROMs. As surrogate outcomes tend to overestimate treatment effects compared with patient-relevant outcomes [36,37], this might also explain why the PRPs had experiences of their assessments' being less valuable to clinicians. Furthermore, the PRPs recommended that questions included in PROMs should be simple and that patients be given the opportunity to explain in words how they truly feel, as questions and scores were perceived as meaningless to them if they felt the content of the PROM was not related to their individual conditions. Such meaninglessness made them question whether PROMs reflect how patients truly feel. This distrust may demonstrate the importance of adding individual clinical experiences of patient care to evidence-based approaches in clinical practice [38], ensuring that patients feel their perspectives are listened to. Nevertheless, Campbell et al. compared questionnaires and interview-based patient assessments of pain and disability after PT treatment for knee-osteoarthritis and found discrepancies (e.g., some patients had worse scores using questionnaires but considered themselves better when interviewed) [39]. This finding may indicate discrepancies in whether patients feel their perspective is attended to and whether clinicians feel they have listened to their patients' perspective. When evaluating treatment effects, different perspectives (e.g., multiple stakeholders) might be involved. Ultimately, however, which is more important: for the clinician or the patient to think there has been improvement?

Our study has limitations. The PROMs included in our study represent a wide variety of outcome measurements tools of which some are well-validated tools, whereas others are undefined or nonvalidated tools. We have not assessed the psychometric characteristics of the questionnaires that have been included in our study to be sure that each instrument is a properly validated tool to capture patients' viewpoint. On this basis, we cannot be confident to what extent the validity, reliability, and responsiveness of each included PROM tool could have influenced our results. Furthermore, the selected outcomes might not have been core outcomes or the most relevant outcomes for each condition and/or intervention. In addition, our main analysis includes comparisons not clearly reflecting the same construct; our subgroup analysis of trials in which comparisons apparently were assessing the same construct accounts for this. However, the influence of patients' and clinicians' different perspectives remained a challenge when interpreting our results. Furthermore, we only included PT-relevant interventions, which may not be representative of other fields. Finally, we extracted data from meta-analyses in the included reviews, and we captured authors' assessments of RoB. To account for improper reporting and lower quality of trials and reviews in PT [40,41], we stratified trials included in reviews published before and after the introduction of the Cochrane RoB tool. Still, we cannot be confident that other metaconfounding did not occur.

In conclusion, although the ROR was compatible with differences in OR of 12%, we found no statistically significant difference between estimates of treatment effects based on PROMs and non-PROMs in trials of PT. When outcomes reflected the same construct, PROMs appeared less favorable than comparable non-PROMs. However, PROMs appeared more favorable compared with *clearly objective* non-PROMs and less favorable when compared with other less objective non-PROMs. The high heterogeneity among trials included in this study indicates the need for further research on which outcome measures are most efficient when assessing treatment effects in various conditions. Patients and clinicians may have different perspectives on treatment effects, and including other instruments/measures together with PROMs should be considered when developing core outcome measurement sets in various conditions.

CRediT authorship contribution statement

Dorthe B. Berthelsen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing - original draft, Writing - review & editing. **Elisabeth Ginnerup-Nielsen:** Investigation, Writing - review & editing. **Carsten Juhl:** Formal analysis, Supervision, Writing - review & editing. **Hans Lund:** Conceptualization, Supervision, Writing - review & editing. **Marius Henriksen:** Writing - review & editing. **Asbjørn Hróbjartsson:** Writing - review & editing. **Sabrina M. Nielsen:** Writing - review & editing. **Marieke Voshaar:** Writing - review & editing. **Robin Christensen:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

Acknowledgments

The authors thank PRPs Solvejg Bager, Jens Christensen, and Anette Jørgensen for their help in preparing the study and discussing the results. During the protocol development, the PRPs participated in discussions of relevance, content, and ethics. Each PRP wrote a hypothesis of expected results, and all these three hypotheses were in agreement with the original hypothesis. During subsequent discussions of results, the PRPs gave their perspective on findings based on their experiences of pros and cons when assessing treatment effects using PROMs and other measurements, respectively. We are grateful to the people who contributed reviews and comments on a previous version of the manuscript; the reviewers their feedback and suggestions for clarity on wording has clearly improved the manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.10.016.

References

[1] Guidance for Industry - Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims: U.S.

Department of Health and Human Services - Food and Drug Administration. 2009. Available at https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf. Accessed September 6, 2017.

[2] Moustgaard H, Bello S, Miller FG, Hrobjartsson A. Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often absent from trial reports. J Clin Epidemiol 2014;67:1327–34.

[3] Hrobjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012;344:e1119.

[4] Hrobjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ 2013;185(4):E201–11.

[5] Hrobjartsson A, Kaptchuk TJ, Miller FG. Placebo effect studies are susceptible to response bias and to other types of biases. J Clin Epidemiol 2011;64:1223–9.

[6] Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol 2014;43:1272–83.

[7] Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601–5.

[8] Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157:429–38.

[9] Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One 2016;11:e0159267.

[10] Khanna G, Singh JA, Pomeroy DL, Gioe TJ. Comparison of patient-reported and clinician-assessed outcomes following total knee arthroplasty. J Bone Joint Surg Am 2011;93. e117(1)-(7).

[11] Efficace F, Rosti G, Aaronson N, Cottone F, Angelucci E, Molica S, et al. Patient- versus physician-reporting of symptoms and health status in chronic myeloid leukemia. Haematologica 2014;99(4):788–93.

[12] Yen JC, Abrahamowicz M, Dobkin PL, Clarke AE, Battista RN, Fortin PR. Determinants of discordance between patients and physicians in their assessment of lupus disease activity. J Rheumatol 2003;30:1967–76.

[13] Evangelou E, Tsianos G, Ioannidis JP. Doctors' versus patients' global assessments of treatment effectiveness: empirical survey of diverse treatments in clinical trials. BMJ 2008;336:1287–90.

[14] Cohen SB, Strand V, Aguilar D, Ofman JJ. Patient- versus physician-reported outcomes in rheumatoid arthritis patients treated with recombinant interleukin-1 receptor antagonist (anakinra) therapy. Rheumatology (Oxford) 2004;43(6):704–11.

[15] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395–400.

[16] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol 2011;64:1303–10.

[17] Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET Handbook: version 1.0. Trials 2017;18(Suppl 3):280.

[18] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745–53.

[19] Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. J Clin Epidemiol 2004;57:543–50.

[20] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65–94.

[21] de Wit MP, Berlo SE, Aanerud GJ, Aletaha D, Bijlsma JW, Croucher L, et al. European League against Rheumatism recommendations for the inclusion of patient representatives in scientific projects. Ann Rheum Dis 2011;70(5):722–6.

[22] Ginnerup-Nielsen E, Christensen R, Thorborg K, Tarp S, Henriksen M. Physiotherapy for pain: a meta-epidemiological study of randomised trials. Br J Sports Med 2016;50:965–71.

[23] Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[24] Higgins JP, Green S. Cochrane Handbook for systematic reviews of interventions. [Chapter 8]. Chichester, Uk: John Wiley & Sons; 2008.

[25] Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med 2000;19:3127–31.

[26] Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21:1513–24.

[27] Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21:1559–73.

[28] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.

[29] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.

[30] WHO. International Statistical Classification of Diseases and Related Health Problems 10th Revision 2016. Avaliable at: http://apps.who.int/classifications/icd10/browse/2016/en. Accessed October 28, 2016.

[31] World Confederation for Physical Therapy. Policy statement: Description of physical therapy 2011. Avaliable at: https://www.wcpt.org/sites/wcpt.org/files/files/PS_Description_PT_Sept2011_FORMATTED_edit2013.pdf. Accessed November 20, 2016.

[32] Guide to Physical Therapist Practice. 2nd ed. American Therapy Association. Physical therapy. Phys Ther 2001;81:9–746.

[33] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.

[34] Basch E, Jia X, Heller G, Barz A, Sit L, Fruscione M, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. J Natl Cancer Inst 2009;101:1624–32.

[35] Mukesh MB, Qian W, Wah Hak CC, Wilkinson JS, Barnett GC, Moody AM, et al. The Cambridge breast intensity-modulated radiotherapy trial: comparison of clinician- versus patient-reported outcomes. Clin Oncol (R Coll Radiol) 2016;28(6):354–64.

[36] Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ 2013;346:f457.

[37] Ridker PM, Torres J. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000-2005. JAMA 2006;295:2270–4.

[38] Ioannidis JP. Evidence-based medicine has been hijacked: a report to David Sackett. J Clin Epidemiol 2016;73:82–6.

[39] Campbell R, Quilty B, Dieppe P. Discrepancies between patients' assessments of outcome: qualitative study nested within a randomised controlled trial. BMJ 2003;326:252–3.

[40] Moseley AM, Herbert RD, Maher CG, Sherrington C, Elkins MR. Reported quality of randomized controlled trials of physiotherapy interventions has improved over time. J Clin Epidemiol 2011;64:594–601.

[41] Moseley AM, Elkins MR, Herbert RD, Maher CG, Sherrington C. Cochrane reviews used more rigorous methods than non-Cochrane reviews: survey of systematic reviews in physiotherapy. J Clin Epidemiol 2009;62:1021–30.