Korean J Pain 2021;34(2):139-155
https://doi.org/10.3344/kjp.2021.34.2.139
pISSN 2005-9159 eISSN 2093-0569

KJP

Review Article

Check for updates

The quality of subgroup analyses in chronic pain randomized controlled trials: a methodological review

Mahmood AminiLari[1,2], Vahid Ashoorian[2], Alexa Caldwell[3], Yasir Rahman[1,2], Robby Nieuwlaat[1], Jason W. Busse[1,2,5], and Lawrence Mbuagbaw[1,4]

[1]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*
[2]*Michael G. DeGroote Institute for Pain Research and Care, McMaster University, Hamilton, Ontario, Canada*
[3]*Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada*
[4]*Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada*
[5]*Department of Anesthesia, McMaster University, Hamilton, Ontario, Canada*

Received September 28, 2020
Revised November 16, 2020
Accepted November 25, 2020

Handling Editor: Hyun Kang

Correspondence
Mahmood AminiLari
Department of Health Research Methods, Evidence and Impact (HEI), Faculty of Health Sciences, McMaster University, 1280 Main Street West, 2C Area, Hamilton, ON L8S 4K1, Canada

Tel: +1-905-554-7283
Fax: +1-905-526-1289
E-mail: aminilam@mcmaster.ca

The quality of subgroup analyses (SGAs) in chronic non-cancer pain trials is uncertain. The purpose of this study was to address this issue. We conducted a comprehensive search in MEDLINE and EMBASE from January 2012 to September 2018 to identify eligible trials. Two pairs of reviewers assessed the quality of the SGAs and the credibility of subgroup claims using the 10 criteria developed by Sun et al. in 2012. The associations between the quality of the SGAs and the studies' characteristics including risk of bias, funding sources, sample size, and the latest impact factor, were assessed using multivariable logistic regression. Our search retrieved 3,401 articles of which 66 were eligible. The total number of SGAs was 177 of which 52 (29.4%) made a subgroup claim. Of these, only 15 (8.5%) were evaluated as being of high quality. Among the 30 SGAs that claimed subgroup effects using an appropriate method of performing interaction tests, the credibility of only 5 were assessed as high. None of the subgroup claims met all the credibility criteria. No significant association was found between the quality of SGAs and the studies' characteristics. The quality of the SGAs performed in chronic pain trials was poor. To enhance the quality of SGAs, scholars should consider the developed criteria when designing and conducting trials, particularly those which need to be specified *a priori*.

Key Words: Bias; Chronic Pain; Logistic Models; MEDLINE; Methods; Pain; Research Design; Uncertainty.

INTRODUCTION

Chronic non-cancer pain (CNCP) refers to pain not due to cancer lasting more than three months [1]. CNCP is a disabling health condition which is highly prevalent and affects approximately 28% of people globally [2]. Randomized controlled trials (RCTs) aim to provide reliable evidence on the efficacy and adverse effects of interventions in general patient populations [3]. However, clinical decisions often depend on individual patient characteristics. Those conducting trials often perform subgroup analyses (SGAs), defined as evaluating the treatment effects in specific subgroups of patients or interventions, to indicate whether the observed treatment effect is altered by base-

Author contributions: Mahmood AminiLari: Methodology; Vahid Ashoorian: Investigation; Alexa Caldwell: Investigation; Yasir Rahman: Data curation; Robby Nieuwlaat: Supervision; Jason W. Busse: Proposal preparation, Analysis plan; Lawrence Mbuagbaw: Supervision.

line characteristics of the study population [4,5]. SGAs thus play a significant role in suggesting the appropriateness of an intervention for a specific patient population and addresses the clinical need for individually based guidelines. They can also inform future studies by determining whether specific baseline prognostic factors may impact outcome measures of interest. However, the practical potential of SGAs can only be realized if an SGA is rigorous in its design and interpretation, as its results may be misleading if incorrectly performed [6].

Numerous criteria have been developed to evaluate the quality of SGAs. Firstly, it is necessary to evaluate if the treatment effect varies across subgroup categories. Since appropriate statistical tests can only identify the extent to which chance explains a study's results and not other factors, performing SGAs without testing for interactions is not a valid technique. More importantly, the lack of *a priori* subgroup hypotheses, and the direction of these interactions can inflate type I error by allowing for multiple hypotheses testing and enhancing the chance of producing spurious subgroup effects [6,7].

Within the literature, it has been found that subgroup claims are often subsequently shown to be incorrect, and that the credibility of subgroup effects is usually low [4]. Notably, a methodological review conducted in the field of chronic back pain found the credibility of subgroup claims to be low [8].

Within the CNCP field, many RCTs have performed SGAs to assess the treatment effects across different subgroups. However, the quality of these analyses and the credibility of the claimed subgroup effects are relatively unknown [8]. There are explicit criteria to help determine the credibility of subgroup effects [4,9,10]. Applying these criteria to CNCP trials, that report SGAs, can help inform the quality of SGAs in this field.

As such, the primary objective of this review was to describe the quality and the credibility of the SGAs conducted in CNCP trials through evaluating their satisfaction of the criteria developed by Sun et al. [4] for assessing the validity of SGAs. Our secondary objective was to explore the associations between studies' characteristics, including risk of bias, funding sources, sample size, and the latest impact factor with the quality of SGAs.

MATERIALS AND METHODS

1. Inclusion criteria

In this study, we included RCTs that were carried out in humans for the management of CNCP. We did not apply restrictions on the basis of study design (parallel, crossover, factorial), number of trial arms, unit of randomization, type of study, study sample size, or category of outcome. To meet inclusion criteria, the RCTs needed to have included one or more SGAs, with or without a subgroup claim. Conference abstracts and publications which were not in English were excluded. The included studies were indexed in MEDLINE and EMBASE from January 2012 to September 2018.

2. Search strategy

An extensive and predefined search strategy (**Appendix 1**) of MEDLINE and EMBASE was conducted from January 2012 to September 2018, using the OVID platform. The strategy's search terms included both MeSH headings and free texts for "subgroup analysis", "chronic pain", " neuropathic pain ", "intervention", "treatment", "management", and "randomized controlled trials".

3. Selection of the eligible studies

Two reviewers (MA and VA), independently and in duplicate, screened titles, and abstracts in the field of pain management to detect citations that were RCTs in humans that performed at least one SGA. For the purposes of this study, we defined an SGA as a statistical analysis that explored whether the effects of an intervention differed according to a sub-group variable. Subsequently, the reviewers, independently and in duplicate, screened the full text of all potentially eligible trials to determine if they met the study's inclusion criteria such as reporting at least one SGA, claiming a subgroup effect using an interaction test, reporting a *P* value for a subgroup effect, and the magnitude of difference in the effect between patient subgroups.

4. Data extraction and management

The data extraction form was created and developed by the principal investigator. At the stage of full text screening, the principal investigator, along with two other reviewers trained in research methodology (MA&VA-MA&YR), extracted information independently and in duplicate from the eligible RCTs. The extracted data included 1) the year of publication, 2) the funding sources, 3) the journal name and latest impact factor (mostly the Thomson Reuters Impact Factor), 4) the trial design, 5) the trial type, 6) the type of participants, 7) the type of intervention and its comparator, 8) the primary outcome(s) and secondary outcome(s), 9) the follow-up duration, 10) the sample size, and 11) the treatment effect for the primary outcome prior to performing the SGA. In the studies that were published as post-hoc analyses of trials, we used additional resources cited

in the included studies, such as published or registered protocols and main trials, to make a more rigorous judgment regarding the quality of the SGAs and the risk of bias assessments.

5. Quality of SGAs

Two pairs of reviewers recorded the number of SGAs performed in each RCT. We assessed the quality and credibility of the SGAs reported using the 10 criteria mentioned above [4]. We assessed the quality of SGAs when the trial performed an SGA but concluded a negative result, and when the trial performed an SGA using an interaction test and claimed a subgroup effect. Due to the various conditions encountered, the following guidelines were developed for the number of criteria considered to evaluate the SGAs:

1) When the trial performed an interaction test and the result was positive (subgroup effect was reported or claimed), all 10 criteria were assessed (credibility).
2) When the trial performed an interaction test and the result was negative (no subgroup effect claimed), 6 criteria were assessed (criteria # 1 to #5 and #7 were applicable).
3) When the trial did not perform an interaction test but reported a positive result (subgroup effect was reported, or the authors reported that the effect appeared larger in one subgroup than another, but acknowledged the fact that they didn't have the power to detect an interaction effect, and therefore these results were considered to be hypothesis generating), 8 criteria were assessed (criteria #5 and #6 were not applicable).
4) When the trial did not perform an interaction test and reported a negative result (no subgroup effect), only the first 4 criteria were assessed.

It should be noted that the first item reflects "credibility", and the next three items reflect the "quality" of SGAs. The quality of all SGAs reported in each study was coded based on the detailed instructions established by Sun et al. [4], which were used in previous studies (**Appendix 2**). Each criterion was scored as 1 if the answer to the item was "yes" (criterion met) and 0 if the answer was "no" (criterion not met). We only assessed the SGA for the pain-related primary outcome and the last follow-up time. If pain was not the primary outcome, we considered the SGA for the primary outcome in addition to the SGA for the most relevant outcome to pain among the secondary outcomes.

Depending on the number of criteria assessed, we scored each SGA between 0 to 10, 0 to 8, 0 to 6, or 0 to 4. We conventionally classified the quality of each SGA based on the proportion of criteria met as high-quality (60% or more) or low quality (less than 60%).

We specifically assessed the credibility of SGAs for those studies which claimed a subgroup effect after performing an interaction test.

6. Risk of bias

Reviewers assessed the risk of bias for included RCTs, independently and in duplicate, using a modified Cochrane risk of bias instrument [11,12]. All disagreements in different stages were resolved by reaching a consensus or consulting with a third reviewer (LM).

7. Data analysis

We used descriptive statistics to summarize and calculate the proportion of trials reporting at least one SGA or claiming a subgroup effect. We also calculated the proportion of SGAs (those which claimed a subgroup effect) meeting each credibility criterion and the number of criteria met by each SGA.

The normality and homogeneity of variance assumptions for continuous outcomes (*e.g.*, functional scores) was verified using the Shapiro–Wilk test and Levene's test, respectively. We performed multivariable linear regression models to assess the potential associations between the quality of the SGAs (as a continues variable) and prespecified study characteristics including the risk of bias (low-risk *vs.* high-risk based on the overall judgment of the reviewers), funding sources (industry and non-industry), sample size (small *vs.* large), and the latest impact factor (as a continues variable). A theory-driven approach was used to build the final multivariable regression model and select the most influential predictor variables. We dichotomized the studies' sample sizes based on the median of this variable into two groups: above and below the median.

To control for the impact of potential multicollinearity issues between the covariates, we calculated the variance inflation factor (VIF) of all variables included in the final models. A VIF of 10 or above (a tolerance of 0.1) was considered as multicollinearity.

To run the regression models, since some of the studies had performed more than one SGA with the same approach to analyzing subgroup effects, we included only one SGA with the highest score in the quality assessment from each study in the regression model. Through applying this approach we limited our analysis to including 66 SGAs, which was equal to the number of studies included. The goodness of fit for the models was also evaluated using the Hosmer–Lemeshow test [13]. Agreement between reviewers regarding: 1) the quality of SGAs, 2) the use of

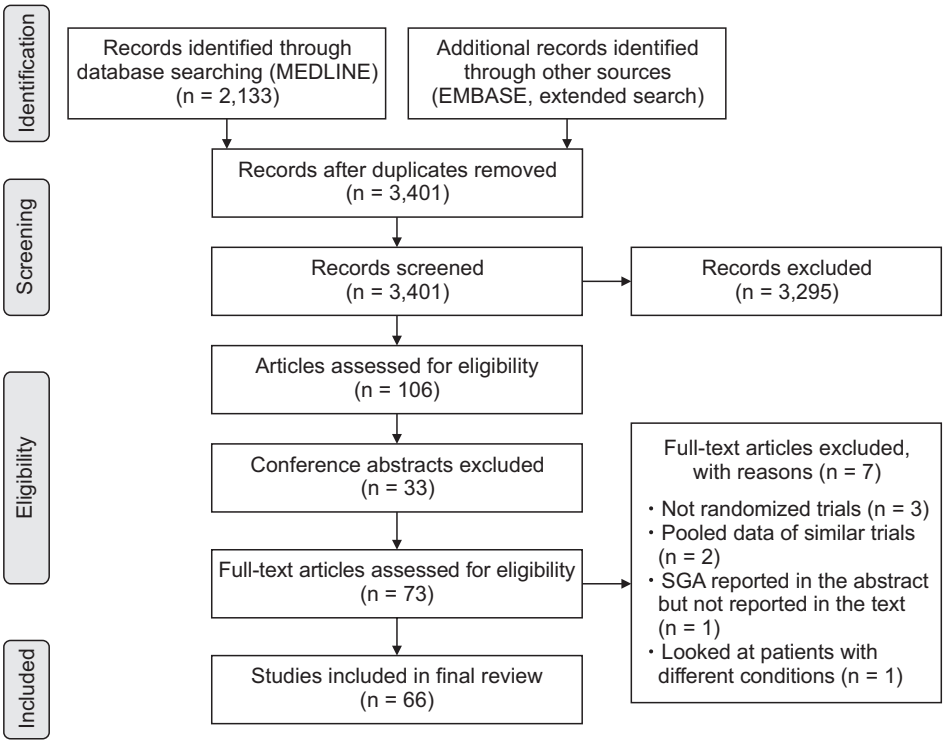

Fig. 1. Study flow diagram. SGA: subgroup analysis.

the interaction test, and 3) the risk of bias assessment was calculated using the Cohen's Kappa statistic. We considered the kappa values of 0-0.20, 0.21-0.40, 0.41-0.60, and 0.61-0.80 as indicating slight, fair, moderate, and substantial agreement, respectively. Values of more than 0.80 were regarded as almost perfect agreement [14]. All analyses were performed using SPSS software version 24 (IBM Co., Armonk, NY).

8. Sample size

To perform the linear regression analysis, we calculated the total number of RCTs that would need to be included. According to Harris and Quade [15], as the rule of thumb for multivariable linear regression analyses, for five or less predictors, the number of subjects should exceed the number of independent variables by 50. For equations involving six or more predictors, an absolute number of 10 subjects per predictor is recommended. Based on these recommendations, a total sample size of at least 60 RCTs was calculated to be included in this study. Considering 4 independent variables for running linear regression models, this study, with 66 RCTs, has sufficient power to produce reliable results.

RESULTS

Two reviewers screened 3,401 titles and abstracts. Of these, 106 publications were potentially identified as eligible. However, 33 articles were conference abstracts, and were thus excluded (**Fig. 1**). The full texts of the remaining 73 studies were retrieved and screened. Sixty-six RCTs were included in the final review, based on the study's eligibility criteria. The descriptions of included studies are reported in **Table 1**, **Appendix 3**.

Table 1. Characteristics of 66 included studies

Study characteristic	Category	Frequency
Trial type	Single center	30 (45.5)
	Multi-center	36 (54.5)
Source of funding	Industry	37 (56.1)
	Non-industry	25 (37.9)
	Both	1 (1.5)
	Not reported	3 (4.5)
Primary outcome (pain)	Yes	43 (65.2)
	No	23 (34.8)
Post-hoc analysis	Yes	37 (56.1)
	No	29 (43.9)
Treatment effect of primary outcome (main trial)	Positive	24 (36.4)
	Negative	42 (63.6)
Risk of bias	High[a]	38 (57.6)
	Low[b]	28 (42.4)

Values are presented as number (%).
[a]High risk of bias: risk of bias evaluated as "High" independently and in duplicate, using a modified Cochrane risk of bias instrument, [b]Low risk of bias: risk of bias evaluated as "Low" independently and in duplicate, using a modified Cochrane risk of bias instrument.

https://doi.org/10.3344/kjp.2021.34.2.139

The inter-rater agreements (Kappa values) for the assessment of the quality of SGAs, the determinant of subgroup claims, and the risk of bias assessment were 0.72 (95% confidence interval [CI]: 0.57-0.87), 0.76 (95% CI: 0.60-0.92), and 0.70 (95% CI: 0.51-0.89), respectively, representing substantial agreement.

Thirty seven out of 66 studies (56.1%) were industry-funded, and 36 (54.5%) were multi-center trials. Within the 66 included studies in the final review, the total number of SGAs reported was 177 (range = 51), and 68.8% of the included studies performed only one SGA. Of these, 52 (29.4%) claimed a subgroup effect. Thirty-two studies (48.5%) performed SGAs using a statistical test for interaction, and the remaining 34 studies (51.5%) performed statistical tests within individual subgroups and compared the results without an interaction test. The frequency of the SGAs, based on the performance of an interaction test (yes or no), is presented in **Table 2**. Among all SGAs, the quality of only 15 (8.5%) was evaluated as high (score ≥ 6 out of 10), and none of the SGAs met all the credibility criteria.

Table 2 also presents the frequency of the SGAs that reported subgroup interactions, which were either positive or negative. Among the 30 (16.9%) SGAs that reported positive results (claimed subgroup effects) using an appropriate method of performing interaction tests, the credibility of only 5 of these SGAs was assessed as high.

Table 3 further indicates the proportion of the above-mentioned 30 SGAs that met each credibility criteria. In 3 SGAs, the subgroup variable was not a characteristic measured at baseline. Additionally, only 1 SGA reported the subgroup variable as a stratification factor at randomization, and only 11 SGAs clearly indicated an *a priori* hypothesis regarding a subgroup effect. Of the 30 claims, only 5 (16.7%) correctly pre-specified the direction of the subgroup effect.

Table 2. Frequency of SGAs categorized based on the result, and performing interaction test[a]

Test of interaction (yes or no)/SGA result (positive or negative)	Frequency	Quality of SGAs	Frequency
Yes/Positive	30 (16.9)	High	5 (16.7)
		Low	25 (83.3)
Yes/Negative	96 (54.2)	High	3 (3.1)
		Low	93 (96.9)
No/Positive	22 (12.4)	High	1 (4.5)
		Low	21 (95.5)
No/Negative	29 (16.4)	High	6 (20.7)
		Low	23 (79.3)

Values are presented as number (%).
SGAs: subgroup analyses.
[a]The 117 SGAs are categorized based on performing an interaction test (yes or no) and the SGA result (positive or negative).

1. Statistical analyses

1) Regression analyses of study variables

We did not find any significant associations using univariate and multivariable regression analyses evaluating the association between the quality of SGAs and the study characteristics (risk of bias, funding sources, sample size, and latest impact factor). The summary of the analyses is presented in **Table 4**.

We assessed the goodness of fit for the final model using the Homer and Lemeshow test. The statistical analysis showed that the Chi-square of 2.241 with 8 degrees of freedom was not significant (P value = 0.973). Therefore, the null hypothesis (H0: The model is appropriate) was rejected and this indicated that the model is appropriate.

Table 3. Proportion of 30 subgroup analyses claiming a subgroup effect which met each criterion

Criteria	No (criterion not met)	Yes (criterion met)
1. Is the subgroup variable a characteristic measured at baseline?	3 (10.0)	27 (90.0)
2. Was the subgroup variable a stratification factor at randomisation?	29 (96.7)	1 (3.3)
3. Was the hypothesis specified *a priori*?	19 (63.3)	11 (36.7)
4. Was the subgroup analysis one of small number of subgroup hypotheses tested (≤ 5)?	10 (33.3)	20 (66.7)
5. Was the test of interaction significant (interaction $P < 0.05$)?	0	30 (100)
6. Was the significant interaction effect independent, if there were multiple significant interactions?	16 (53.3)	14 (46.7)
7. Was the direction of subgroup effect correctly pre-specified?	25 (83.3)	5 (16.7)
8. Was the subgroup effect consistent with evidence from previous studies?	20 (66.7)	10 (33.3)
9. Was the subgroup effect consistent across related outcomes?	20 (66.7)	10 (33.3)
10. Was there indirect evidence to support the apparent subgroup effect (biological rationale, laboratory tests, animal studies)?	28 (93.3)	2 (6.7)

Values are presented as number (%).

Table 4. Association between quality of SGAs with studies' characteristics using multiple linear regression models

Variable	Univariable analysis		Multivariable analysis	
	B (95% CI)	*P* value	B (95% CI)	*P* value
Risk of bias	0.33 (−0.24, 0.91)	0.258	0.16 (−0.45, 0.78)	0.591
Source of funding	−0.005 (−0.61, 0.60)	0.986	−0.05 (−0.71, 0.61)	0.880
Sample size	0.15 (−0.41, 0.73)	0.586	−1.81 (−0.99, 0.63)	0.658
Journal impact factor	0.33 (−0.24, 0.91)	0.258	0.23 (−0.39, 0.85)	0.461

SGAs: subgroup analyses, CI: confidence interval.

DISCUSSION

1. Summary and interpretation of findings

In this methodological study, we assessed the quality and credibility of SGAs performed in CNCP trials published between 2012 and 2018. SGAs aim to detect a subset of the patient population with improved efficacy when compared to the whole trial population, based on specific patients or intervention characteristics. Of the 66 included studies that reported at least one SGA, a higher proportion of the included studies was industry-funded, indicating that a higher proportion of industry-funded trials reported an SGA compared to non-industry funded trials.

Another variable influencing the quality of SGAs is sample size. Lachenbruch [16] suggested a simple method of calculating a trial's sample size for it to be eligible to test for subgroup interactions using the contrast(s) for the interaction and a normal distribution. A required sample size of approximately 500 has also suggested by previous studies [17]. Based on these two rationales, 79% of the included studies did not meet the requirements and were considerably underpowered to detect any significant subgroup effects. This issue highlights the lack of power for performing SGAs.

The quality of SGAs is also influenced by the number of the subgroup hypotheses that were tested. In this study, approximately two-thirds of the included studies performed only one SGA and 7.5% of the studies performed more than 5 SGAs, leading them to exceed the quality criterion that less than 5 subgroup hypotheses should be tested. Performing many interaction tests in one study could suggest a significant inflation of type I error, which could enhance the probability of reporting spurious results.

Additionally, in slightly less than 50% of the studies, the authors expressed that they undertook an interaction test for analyzing subgroups, and reported a *P* value for interaction. A test for interaction, which examines if the treatment effect varies across subgroup categories, is the only reliable statistical approach to claim that the existing difference between subgroups cannot be explained by chance [10,18].

Overall, the quality of SGAs performed in the 66 included studies was low. Among the 177 SGAs identified, the quality of only 15 (8.5%) was high. Of the 30 SGAs that claimed a subgroup effect using an appropriate test for interaction, the credibility of only 5 SGAs was evaluated as high. According to **Table 3**, approximately two-thirds of the SGAs claiming a subgroup effect failed to clearly indicate an *a priori* hypothesis for the subgroup effect. Even when subgroup effects were hypothesized *a priori*, the direction of a majority of subgroup effects (83%) was not correctly hypothesized *a priori*. One reason for this could be about that 56% of the included studies were post-hoc analyses of RCTs. This result may be explained by the fact that these SGAs were carried out to find significant differences in primary outcome measures in specific patient subgroups when one was not found in an analysis of the whole study population. However, this study did not correlate this parameter of SGA quality with the primary outcome results for the whole study populations of the 44% of studies that did not generate *a priori* hypotheses. As such, this remains a hypothesis that warrants further study.

Nevertheless, of the studies which performed a test for interaction between subgroups, 90% of them satisfied this criterion that "the subgroup variable was a characteristic that was measured at baseline". This indicates that most of the SGAs were selected based on characteristics at baseline.

Overall, the results of this study indicate that a total of 52 SGAs reported a subgroup effect. However, in 22 of these subgroup effects, the authors concluded that there was a subgroup effect by reporting a significant treatment effect in one subgroup or by looking for significance in each subgroup separately which cannot be considered as a correct method of claiming a subgroup effect [18].

Independence of the interaction is an important criterion whose fulfillment in performing SGAs can increase the credibility of subgroup effects. When a study tests multiple hypotheses, the analyses might produce more than one significant interaction which might be associated with each other and explained by a common factor [10]. This is-

sue can be addressed by including all significant and non-significant interactions in the regression model to see if the interaction terms are still significant. In our study, of the 30 claims, 14 (46.7%) met this criterion by performing regression models to check if the interaction term was independent.

2. Strengths and limitations of the study

To our knowledge, the current study is the first methodological review conducted to assess the quality of SGAs among all non-cancer chronic pain trials after the publication of the 10 criteria to assess SGA validity in 2012 [4]. There is just one similar review [8]; however, our study differs in two important regards. Firstly, our study evaluated the quality of SGAs reported in all non-cancer chronic pain trials while the scope of the previous review was narrower and included specifically low back pain trials with SGAs. Secondly, our study assessed the quality and the credibility of all SGAs reported (positive and negative) rather than just looking at those with a claim of a subgroup effect. As such, we deem our review of the literature to be more robust.

Furthermore, given the variety of studies with different forms of SGAs, we divided the SGAs into 4 categories based on the test of interaction performed and the result of the SGAs (positive-negative) and evaluated the quality or credibility of each subgroup based on the number of criteria applied in each category. The previously available tools were designed to assess the credibility of subgroup effects claimed in the RCTs; however, there was no standard tool to take into consideration the quality of performing all SGAs rather than only those which reported a claim. As such, our approach allowed for a more stratified and appropriate evaluation of the SGAs performed.

Our study is also presented with two limitations. Firstly, based on the initial study protocol, we searched MEDLINE starting with 2013. Due to not obtaining the required sample size (60), we expanded our search to EMBASE and to the year 2012 to obtain more eligible studies. Since we limited the literature search to studies published in or after 2012 to coincide with the publication of the guidelines created by Sun et al. [4] and for it to thus have been possible for the SGAs to have been designed in accordance to those guidelines, we were only able to include 66 RCTs.

The results of our study are consistent with the findings of previous studies conducted on this issue [8]. Previous searches of the literature have also demonstrated the poor quality of SGAs and the low credibility of subgroup claims.

Contrary to what we expected, no significant association was found between the quality of SGAs, and the risk of bias, the source of funding, the sample size, or the journal impact factor. This finding indicates that the quality of SGAs might not be affected by study characteristics. One reason for this could be the small sample size which might have made our study underpowered to reach actual associations between study variables. Other studies have also reported a lack of association between study characteristics and SGA quality [8,17]. However, the source of funding was not a study characteristic included in the previous multivariable regressions published in the literature.

The results of the current study, in keeping with the results of previous studies [19,20] show that a larger proportion of included trials were funded by industry. It is possible that this result indicates that, in the presence of non-significant results (73% *vs.* 27% in our study), industry funded trials may be more likely to attempt to seek statistically significant findings in patient subgroups. However, our multiple regression analyses did not prove this claim.

3. Conclusion

The findings of this study indicated that the overall quality of SGAs and the credibility of subgroup effects in CNCP trials is low. This study emphasizes the importance of utilizing appropriate scientific methodology to investigate subgroup effects and highlights the following issues: Those conducting trials should utilize the standardized criteria, specifically in the process of trial planning. Utilizing experienced statisticians to include SGAs in the analyses planning is highly recommended. Journal editors should also consider the developed criteria to assess the credibility of subgroup claims reported in the submitted manuscripts. Finally, knowledge users should also take caution in their interpretation of the results of SGAs and their application of the treatment in question to specific subpopulations.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

No funding to declare.

ORCID

Mahmood AminiLari, https://orcid.org/0000-0002-0517-9666
Vahid Ashoorian, https://orcid.org/0000-0001-9225-155X
Alexa Caldwell, https://orcid.org/0000-0002-0223-6303

Yasir Rahman, https://orcid.org/0000-0002-2817-3638
Robby Nieuwlaat, https://orcid.org/0000-0002-5267-7748
Jason W. Busse, https://orcid.org/0000-0002-0178-8712
Lawrence Mbuagbaw, https://orcid.org/0000-0001-5855-5461

REFERENCES

1. Ospina M, Harstall C. Prevalence of chronic pain: an overview. Edmonton, Alberta Heritage Foundation for Medical Research. 2002.
2. Elzahaf RA, Tashani OA, Unsworth BA, Johnson MI. The prevalence of chronic pain with an analysis of countries with a Human Development Index less than 0.9: a systematic review without meta-analysis. Curr Med Res Opin 2012; 28: 1221-9.
3. Venekamp RP, Rovers MM, Hoes AW, Knol MJ. Subgroup analysis in randomized controlled trials appeared to be dependent on whether relative or absolute effect measures were used. J Clin Epidemiol 2014; 67: 410-5.
4. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ 2012; 344: e1553.
5. Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. J Biopharm Stat 2014; 24: 154-67.
6. Byth K, Gebski V. Factorial designs: a graphical aid for choosing study designs accounting for interaction. Clin Trials 2004; 1: 315-25.
7. McCormack R, Lamontagne M, Vannabouathong C, Deakon RT, Belzile EL. Comparison of the 3 different injection techniques used in a randomized controlled study evaluating a cross-linked sodium hyaluronate combined with triamcinolone hexacetonide (Cingal) for osteoarthritis of the knee: a subgroup analysis. Clin Med Insights Arthritis Musculoskelet Disord 2017; 10: 1179544117725026.
8. Saragiotto BT, Maher CG, Moseley AM, Yamato TP, Koes BW, Sun X, et al. A systematic review reveals that the credibility of subgroup claims in low back pain trials was low. J Clin Epidemiol 2016; 79: 3-9.
9. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116: 78-84.
10. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010; 340: c117.
11. Oxman A, Guyatt G, Cook D, Montori V. Summarizing the evidence. In: Users' guides to the medical literature: a manual for evidence-based clinical practice. Edited by Guyatt G, Rennie D. Chicago, AMA Press. 2002, pp 155-173.
12. Akl EA, Sun X, Busse JW, Johnston BC, Briel M, Mulla S, et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. J Clin Epidemiol 2012; 65: 262-7.
13. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Commun Stat Theory Methods 1980; 9: 1043-69.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.
15. Harris RJ, Quade D. The minimally important difference significant criterion for sample size. J Educ Stat 1992; 17: 27-49.
16. Lachenbruch PA. A note on sample size computation for testing interactions. Stat Med 1988; 7: 467-9.
17. Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. Spine (Phila Pa 1976) 2014; 39: 618-29.
18. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365: 176-86.
19. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. BMJ 2011; 342: d1569.
20. Barton S, Peckitt C, Sclafani F, Cunningham D, Chau I. The influence of industry sponsorship on the reporting of subgroup analyses within phase III randomised controlled trials in gastrointestinal oncology. Eur J Cancer 2015; 51: 2732-9.

Appendices

Appendix 1. Search strategy

Database: Ovid MEDLINE epub ahead of print, in-process & other non-indexed citations, Ovid MEDLINE(R) daily and Ovid MEDLINE(R) 1946 to present
Search strategy:

1	(chronic adj4 pain*).mp. (60364)
2	chronic pain/ (10418)
3	exp osteoarthritis/ (56304)
4	osteoarthrit*.mp. (78524)
5	osteo-arthritis.mp. (370)
6	degenerative arthrit*.mp. (1235)
7	exp rheumatoid arthritis/ (106226)
8	exp neuralgia/ (18213)
9	diabetic neuropathy/ (13813)
10	(neuropath* adj5 (pain* or diabet*)).mp. (38054)
11	neuralg*.mp. (24465)
12	zoster.mp. (19578)
13	irritable colon/ (6314)
14	(Irritable Bowel Syndrome or IBS).mp. (14170)
15	exp migraine/ (25156)
16	migraine.mp. (35284)
17	fibromyalgia/ (7739)
18	fibromyalg*.mp. (10571)
19	reflex sympathetic dystrophy.mp. (4041)
20	(complex regional pain syndromes or causalgia).mp. (2186)
21	intractable pain/ (6051)
22	phantom limb.mp. or agnosia/ or phantom pain/ or amputation stump/ (7218)
23	hyperalgesia/ (10385)
24	((noncancer* or non-cancer*or chronic* or recurrent or persist* or non-malign*) adj3 pain).mp. (17023)
25	or/1-24 (388401)
26	clinical trial/ (512199)
27	controlled study/ (0)
28	exp clinical study/ (859596)
29	randomized controlled trial/ (467803)
30	double blind procedure/ (0)
31	multicenter study/ (238803)
32	single blind procedure/ (0)
33	phase 3 clinical trial/ (0)
34	phase 4 clinical trial/ (0)
35	crossover procedure/ (0)
36	placebo/ (0)
37	allocat$.mp. (203144)
38	assign$.mp. (282136)
39	blind$.mp. (332143)
40	(clinic$ adj25 (study or trial)).mp. (1300807)
41	compar$.mp. (5670215)
42	control$.mp. (4869672)
43	cross?over.mp. (57020)
44	factorial$.mp. (26363)
45	follow?up.mp. (19893)
46	placebo$.mp. (211883)
47	prospectiv$.mp. (767121)
48	random$.mp. (1217036)
49	((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).mp. (224239)
50	trial.mp. (1081277)

51	(versus or vs).mp. (1170229)
52	or/37-51 (10141777)
53	subgroup analysis.mp. (19173)
54	moderator.mp. (6154)
55	effect modifier.mp. (1963)
56	interaction.mp. (718086)
57	subpopulation.mp. (30080)
58	subset.mp. (140541)
59	or/53-58 (904752)
60	25 and 52 and 59 (6034)
61	limit 60 to humans (4899)
62	limit 61 to yr="2012 -Current" (2133)

Database: Embase <1996 to 2018 August 31>
Search strategy:

1	(chronic adj4 pain*).mp. (90141)
2	chronic pain/ (48559)
3	exp osteoarthritis/ (95666)
4	osteoarthrit*.mp. (106474)
5	osteo-arthritis.mp. (227)
6	degenerative arthrit*.mp. (958)
7	exp rheumatoid arthritis/ (137872)
8	exp neuralgia/ (78995)
9	diabetic neuropathy/ (17277)
10	(neuropath* adj5 (pain* or diabet*)).mp. (59624)
11	neuralg*.mp. (21562)
12	zoster.mp. (25946)
13	irritable colon/ (20767)
14	(Irritable Bowel Syndrome or IBS).mp. (20572)
15	exp migraine/ (47073)
16	migraine.mp. (51808)
17	fibromyalgia/ (16696)
18	fibromyalg*.mp. (17882)
19	reflex sympathetic dystrophy.mp. (1438)
20	(complex regional pain syndromes or causalgia).mp. (797)
21	intractable pain/ (3044)
22	phantom limb.mp. or agnosia/ or phantom pain/ or amputation stump/ (5403)
23	hyperalgesia/ (16796)
24	((noncancer* or non-cancer*or chronic* or recurrent or persist* or non-malign*) adj3 pain).mp. (21836)
25	or/1-24 (535124)
26	clinical trial/ (814030)
27	controlled study/ (5491827)
28	exp clinical study/ (6325695)
29	randomized controlled trial/ (464812)
30	double blind procedure/ (125723)
31	multicenter study/ (184132)
32	single blind procedure/ (30760)
33	phase 3 clinical trial/ (33715)
34	phase 4 clinical trial/ (2822)
35	crossover procedure/ (51817)
36	placebo/ (269065)
37	allocat$.mp. (126145)
38	assign$.mp. (299040)
39	blind$.mp. (355535)
40	(clinic$ adj25 (study or trial)).mp. (4549637)

41 compar$.mp. (5803479)
42 control$.mp. (7892083)
43 cross?over.mp. (73398)
44 factorial$.mp. (52978)
45 follow?up.mp. (35513)
46 placebo$.mp. (338337)
47 prospectiv$.mp. (925030)
48 random$.mp. (1374029)
49 ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).mp. (223919)
50 trial.mp. (1676314)
51 (versus or vs).mp. (1692991)
52 or/37-51 (12375015)
53 subgroup analysis.mp. (30202)
54 moderator.mp. (6812)
55 effect modifier.mp. (1092)
56 subpopulation.mp. (76836)
57 53 or 54 or 55 or 56 (114732)
58 25 and 52 and 57 (3096)
59 limit 58 to human (2699)
60 limit 59 to yr="2012 -Current" (1779)

Appendix 2. Criteria to assess the credibility of subgroup claims [4]

Criterion	Description of criteria	Coding
Design		
1. Is the subgroup variable a characteristic measured at baseline?	Subgroup variables measured after randomisation might be influenced by the tested interventions. The apparent difference of treatment effect between subgroups can be explained by the intervention, or by differing prognostic characteristics in subgroups that appear after randomisation.	Yes, if the study specified that subgroups were defined on the basis of characteristics at baseline. No, if the study describes that the subgroups were defined according to characteristics measured after randomisation or did not describe when the subgroups were defined.
2. Was the subgroup variable a stratification factor at randomisation?	Credibility of subgroup difference would be increased if a subgroup variable was also used for stratification at randomisation (*i.e.*, stratified randomisation).	Yes, if the randomisation included stratification based on the pre-specified subgroups variable. No, if the study clearly reported information on stratification, but the subgroup variable of interest was not one of the stratification factors, or if no information was available regarding stratification.
3. Was the hypothesis specified *a priori*?	A subgroup analysis might be clearly planned before to test a hypothesis. This must be mentioned on the study protocol (registered or published) or primary trial, when appropriate. Post-hoc analyses are more susceptible to bias as well as spurious results and they should be viewed as hypothesis generating rather than hypothesis testing.	Yes, there needs to be a publicly available record (*i.e.*, study protocol, registry, or primary trial) of the hypothesis that predates the subgroup analyses. No, if the report specifically says the analyses were post-hoc, or no information reported regarding this aspect.
4. Was the subgroup analysis one of small number of subgroup hypotheses tested (≤ 5)?	The greater the number of hypotheses tested, the greater the number of interactions that will be discovered by chance, that is, the more likely it is to make a type I error (reject one of the null hypotheses even if all are actually true). A more appropriate analysis would account for the number of subgroups.	Yes, if the authors clearly specify up to 5 subgroup hypotheses. No, if authors clearly specified more than 5 subgroup hypotheses, or if the study did not give this information.
Analysis		
5. Was the test of interaction significant (interaction $P < 0.05$)?	Statistical tests of significance must be used to assess the likelihood that a given interaction might have arisen due to chance alone (the lower a *P* value is, the less likely it is that the interaction can be explained by chance).	Yes, if the study used any reliable statistical test to assess the subgroup interactions (*e.g.*, regression models), and a *P* value lower than 0.05. No, no reliable statistical test used, or *P* value higher than 0.05.
6. Was the significant interaction effect independent, if there were multiple significant interactions?	When testing multiple hypotheses in a single study, the analyses might yield more than one apparently significant interaction. These significant interactions might, however, be associated with each other, and thus explained by a common factor.	Yes, if the significant subgroup effect was not associated with other significant interactions, or if the subgroup effect was tested regarding its independence with other interaction effects (usually tested in multivariable regression that includes interaction terms). No, if the subgroup effect was analysed only as part of a significant interaction effect.

Appendix 2. Continued

Criterion	Description of criteria	Coding
Context		
7. Was the direction of subgroup effect correctly pre-specified?	A subgroup effect consistent with the pre-specified direction will increase the credibility of a subgroup analysis. Failure to specify the direction or even getting the wrong direction weakens the case for a real underlying subgroup effect.	Yes, if the direction of subgroup effect was correctly specified *a priori* (*e.g.*, study protocol, published statistical analysis plan, trial registry). No, if the authors fail to specify the direction or specify the wrong direction *a priori*.
8. Was the subgroup effect consistent with evidence from previous studies?	A hypothesis concerning differential response in a subgroup of patients may be generated by examination of data from a single study. The interaction becomes far more credible if it is also found in other similar studies. The extent to which a comprehensive scientific overview of the relevant literature finds an interaction to be consistently present is probably the best single index as to whether it should be believed. In other words, the replication of an interaction in independent, unbiased studies provides strong support for its believability.	Yes, if the study provides information that there was a consistent interaction found in other studies consistent with both the power of the comparisons and differences between studies that might influence results. No, if the information provided by the study was not consistent across other studies, or if no information about other studies were reported.
9. Was the subgroup effect consistent across related outcomes?	The subgroup effect is more likely to be real if its effect manifest across all closely related outcomes. Studies must determine whether the subgroup effect existed among related outcomes.	Yes, if there was a consistent interaction of a subgroup across closely related outcomes within the study; that is, there was a consistency of the subgroup effect across the related outcomes. No, if the study did not determine whether the subgroup effect exists across the related outcomes.
10. Was there indirect evidence to support the apparent subgroup effect (biological rationale, laboratory tests, animal studies)?	We are generally more ready to believe a hypothesised interaction if indirect evidence makes the interaction more plausible. That is, to the extent that a hypothesis is consistent with our current understanding of the biologic mechanisms of disease, we are more likely to believe it. Such understanding comes from three types of indirect evidence: (i) from studies of different populations (including animal studies); (ii) from observations of interactions for similar interventions; and (iii) from results of studies of other related outcomes.	Yes, if the study provides information that the consistent interaction of a subgroup is plausible to indirect evidence. No, the significant interaction found was not reasonable with indirect evidence, or no information reported regarding this issue.

Appendix 3. The summary of included studies

Study ID	First author	Year	Latest impact factor	Country	Sample size	Chief complaint	Intervention(s)	Comparator	Primary outcome measure	Follow-up duration (days)
1	Geusens	2017	1.99	Belgium	211	Primary knee osteoarthritis (OA)	Choline-stabilized orthosilicic acid	Placebo	Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain	84
2	Cook	2015	3.02	USA	179	Chronic nonspecific neck pain	Massage	No additional massage	Neck pain	182
3	Smelt	2012	6.2	The Netherlands	490	Migraine	Training on treating migraine and participating patients invited for a consultation/evaluation of their therapy	Usual care	Headache Impact Test (HIT-6)	365
4	Garcia	2016	2.58	Brazil	140	Nonspecific low back pain	Mechanical diagnosis and therapy	Back School (a group-based treatment approach)	Pain intensity	30
5	Arden	2014	2.7	UK, Sweden, and Germany	218	Knee pain	A single IA injection in the study knee of either NASHA (Durolane 60 mg in 3 mL)	Saline (3 mL phosphate buffered saline, pH 7)	WOMAC pain	42
6	Takamura	2018	2.62	Japan	809	Knee OA	Gel-200 or phosphate buffered saline	Phosphate buffered saline	WOMAC pain	182
7	Aaron Yarlas	2014	2.18	USA	541	Chronic low back pain (CLBP)	The Butrans (buprenorphine) Transdermal System (BTDS)	Matching placebo patch	Pain severity	7
8	Marchettini	2015	2.9	Multi-Country	804	Diabetic peripheral neuropathic pain (DPNP)	Duloxetine (60 mg/day) or pregabalin (300 mg/day)	Increasing each drug to its maximum dose	Pain intensity	56
9	Castel	2015	1.95	Spain	130	Fibromyalgia	Multidisciplinary treatment: cognitive behavioural therapy (CBT), and physical therapy	Conventional pharmacologic treatment	Pain intensity, functionality, catastrophizing	365
10	Broderick	2016	5.59	USA	256	Chronic pain due to OA of the knee or hip	Pain coping skills training	Usual care	Pain	168
11	Tullo	2012	2.28	Italy	107	Migraine	Frovatriptan	Zolmitriptan 2.5 mg	Pain	2
12	Monfort	2014	3.3	Spain	100	OA	Ultrasound-guided intra-articular treatment with hyaluronic acid	Betamethasone	Clinical improvement by the FIHOA score	180
13	McCormack	2017	No IF	Canada	368	Kellgren-Lawrence (KL) grade I-III knee OA	Cingal	Both Monovisc and saline	WOMAC pain	84
14	Jürg Schliessbach	2018	2.76	Switzerland	50	Chronic low-back pain	Imipramine 75 mg	Active placebo	Intensity of low-back pain	0.08
15	Stamuli	2012	2.73	UK	233	Irritable bowel syndrome	Short course of traditional acupuncture plus usual GP care	Usual GP care	The EQ-5D to measure and value patients' health states	365
16	R. Forestier	2014	No IF	France	214	Knee OA	Crenobalneotherapy (spa) therapy	Exercises and usual treatment alone	WOMAC function subscale	252
17	Duivenvoorden	2015	4.09	The Netherlands	91	Symptomatic medial compartmental knee OA	A laterally wedged insole	Valgus brace	Pain severity	42
18	Rikke K. Jensen	2015	No IF	Denmark	96	Persistent low back pain	A new treatment approach comparing rest	Exercise group	Low back pain intensity	426
19	Annemieke J.H	2015	2.58	The Netherlands	135	Sciatica	Physical therapy plus general practitioners' care	General practitioners' care alone	Leg pain intensity	365
20	Kristen Radcliff	2013	2.79	USA	501	Radicular pain with positive nerve root tension sign	Surgical treatment (lumbar discectomy)	Nonoperative treatment	Bodily pain, physical function, and mental component summary domains of the SF-36	730

https://doi.org/10.3344/kjp.2021.34.2.139

Appendix 3. Continued

Study ID	First author	Year	Latest impact factor	Country	Sample size	Chief complaint	Intervention(s)	Comparator	Primary outcome measure	Follow-up duration (days)
21	Tanenberg	2013	2.18	USA	407	DPNP	The combination of duloxetine, 60 mg/d, and gabapentin, 900 mg/d or more duloxetine monotherapy, 60 mg/d	Pregabalin monotherapy	Daily pain	84
22	Verbruggen	2012	12.35	Belgium	60	Erosive hand OA	40 mg adalimumab	Placebo	Reduction in progression of structural damage	365
23	Sumaiyah Mat	2018	1.85	Malaysia	41	Knee OA	A home-based modified Otago Exercise Program	General health advice and conventional treatment	Fear of falling, OA symptoms, and functional ability	182
24	Ben-Ami	2017	3.09	Israel	220	CLBP	Enhanced transtheoretical model intervention	Physical therapy	Disability	365
25	Bergström	2012	2.04	Sweden	214	Patients with chronic neck pain	Behavioural-oriented physiotherapy (PT), CBT, behavioural medicine rehabilitation (BM)	Treatment-as-usual	Registered sickness absence	3,650
26	Kvalvaag	2017	6.05	Norway	143	Subacromial pain syndrome	Audial extracorporeal shock wave therapy (reswt) in addition to supervised exercises	Sham rest in addition to supervised exercises	Shoulder pain	168
27	Brämberg	2017	1.99	Sweden	159	Non-specific low back pain	Kundalini yoga- strength training	Self-care evidence-based advice	Sickness absenteeism	365
28	Hahne	2017	3.11	Australia	54	Clinical features of radiculopathy and imaging showing a lumbar disc herniation	Individualized functional restoration incorporating advice	Guideline-based advice alone	Activity limitation, back pain and leg pain	364
29	Kellner	2012	2.7	USA	4,484	OA -rheumatoid arthritis (RA)	Celecoxib	Diclofenac SR plus omeprazole	A composite of clinically significant upper and lower gastrointestinal (GI) events	182
30	Ashina	2018	3.88	Denmark	667	Chronic migraine	Monthly erenumab (70 mg or 140 mg)	Placebo	Change in reduced monthly migraine days	90
31	Sarvajeet PAL	2016	2.42	India	3,366	RA	Add-on subcutaneous golimumab therapy	Conventional dmards regimen	Proportion of patients achieving good to moderate DAS28-ESR	182
32	Broglio	2017	2.18	USA	922	Chronic non-malignant and nonneuropathic pain	Hydrocodone bitartrate (HYD)	Placebo	Patient-reported pain intensity	365
33	Yiming MU	2018	7.27	China	620	Painful diabetic peripheral neuropathy	300 mg/day pregabalin	Placebo	Pain	56
34	Tanaka	2018	1.95	Japan	878	RA	Sirukumab subcutaneously	Placebo	The American College of Rheumatology (ACR) 20 response at week 16	112
35	Genovese	2015	1.97	Multi-Country	1,004	RA	Tabalumab every 4 weeks (120/Q4W), 90 mg tabalumab every 2 weeks (90/Q2W)	Placebo	ACR 20 (20% improvement in American College of Rheumatology criteria) response at week 24	168
36	Palacios	2016	2.67	Mexico	7,492	Osteoporosis	Bazedoxifene (BZA) 20 mg/d, BZA 40 mg/d, raloxifene 60 mg/d	Placebo	Serum concentrations of osteocalcin	1,095
37	Mease	2014	2.58	USA	336	Fibromyalgia	Milnacipran (MLN)	Placebo	Pain	365
38	Zheng	2016	2.02	China	448	Irritable bowel syndrome	The electroacupuncture	Loperamide group	Stool frequency	28
39	Bartolini	2012	3.34	Italy	114	Migraine	Frovatriptan 2.5 mg	Almotriptan 12.5 mg	The proportion of pain-relief episodes	2

Appendix 3. Continued

Study ID	First author	Year	Latest impact factor	Country	Sample size	Chief complaint	Intervention(s)	Comparator	Primary outcome measure	Follow-up duration (days)
40	Kanzaki	2015	2.5	Japan	100	Knee pain	1,200 mg of glucosamine hydrochloride, 60 mg of chondroitin sulfate, 45 mg of type II collagen peptides, 90 mg of quercetin glycosides, 10 mg of imidazole peptides, and 5 μg of vitamin D per day	Placebo	Knee-joint functions	112
41	Chan	2017	3.12	Australia	96	Non-reducible discogenic pain (NRDP)	Individualized functional restoration plus guideline-based advice	Advice alone	Disability back and leg pain	182
42	Thackeray	2016	3.09	USA	120	Low back pain and nerve root compression	Physical therapy following an extension-oriented treatment approach with the addition of mechanical traction	Physical therapy without the addition of mechanical traction	Low back pain-related disability	365
43	Bennell	2015	4.14	Australia	100	Medial knee OA	Exercise programs quadriceps strengthening	Neuromuscular exercise (NEXA)	Overall knee pain	365
44	Baratloo	2016	0.8	Iran	110	Migraine headaches	Caffeine versus ketorolac	Ketorolac	Pain intensity	0.08
45	Knoop	2014	1.82	The Netherlands	159	Knee OA and knee instability	Knee stabilization therapy, prior to strength/functional training	Exercise program without knee stabilization therapy	WOMAC subscale physical function	266
46	Haufe	2017	2.76	Germany	226	Low back pain	5-month non-supervised training at home	Wait-list-control	Isometric back extension strength associated with low back pain disability and function	140
47	Garcia	2018	7.86	Brazil	148	CLBP	Mckenzie Method of Mechanical Diagnosis and Therapy (MDT)	Placebo	Pain intensity and disability	365
48	Weegen	2015	3.33	The Netherlands	196	Symptomatic knee OA	A specific brand of hyaluronic acid	Placebo	Range of motion (ROM), pain	182
49	Aurora SK	2014	3.12	USA	1,384	Migraine	Onabotulinumtoxina (155 U)	Placebo	Frequency of headache	168
50	Amerongen	2017	3.18	The Netherlands	24	Progressive primary or secondary MS	Oral formulation of D9-Tetrahydrocannabinol	Placebo	Objective spasticity	28
51	Losina	2016	4.14	USA	308	Total knee arthroplasty (TKA)	Postoperative care with frequent follow-up by a care navigator	Usual care	WOMAC physical function	182
52	Petersen	2015	1.99	Denmark	350	CLBP	The Mckenzie method or MDT	Spinal manipulation treatment	Disability	60
53	Amris	2016	1.8	Denmark	191	Fibromyalgia and chronic widespread pain (CWP)	Group-based multi-component rehabilitation program	Waiting list	Motor and process skills	182
54	Licciardone	2013	Not found	USA	144	Low back pain	Usual obstetric care and osteopathic manipulative treatment (UOBC + OMT), usual obstetric care and sham ultrasound therapy (UOBC + SUT)	UOBC	Progressive back-specific dysfunction	63
55	Cook	2015	3.1	USA	191	Chronic non-specific neck pain	Massage	A wait-list control group	Neck pain	182
56	Licciardone	2016	0.7	USA	455	CLBP	A short regimen of OMT	Sham-controlled	LBP intensity	84
57	Overdevest	2014	3.19	The Netherlands	283	Sciatica due to a lumbar disc herniation	Surgery	Prolonged conservative treatment	Recovery of motor deficit	364
58	Price	2017	3.2	Canada	70	Postherpetic neuralgia	Tv-45070	Placebo ointments	Pain intensity	21

Appendix 3. Continued

Study ID	First author	Year	Latest impact factor	Country	Sample size	Chief complaint	Intervention(s)	Comparator	Primary outcome measure	Follow-up duration (days)
59	Kwon	2013	2.84	USA	300	Pain associated with glenohumeral osteoarthritis (GH-OA)	Sodium hyaluronate (HA)	Phosphate-buffered saline (PBS)	Pain	182
60	Kessler	2014	2.76	Switzerland	60	Chronic pelvic pain syndrome (cpps)	Sono-electro-magnetic therapy	Placebo	A change in the national institutes of health chronic prostatitis symptom index	84
61	Jensen	2015	1.95	USA	25	Chronic pain	Consumption of hydrolyzed water-soluble egg membrane	Placebo	ROM, pain at rest	28
62	Mohs	2012	3.81	USA	154	Fibromyalgia	Duloxetine	Placebo	Speed of processing on tasks requiring visual attention, working memory, and executive function	84
63	McGuire	2014	3.11	USA	749	Lumbar pathology	Operative versus nonoperative treatments	Nonoperative treatments	The components of the SF-36	730
64	Enomoto	2018	1.21	Japan	353	Patients with knee OA pain	Duloxetine	Placebo	Pain	98
65	Schlögl	2018	5.32	Switzerland	200	Chronic pain	Vitamin D	Combination of calcifediol and vitamin D3	Change in the mean number of painful areas	365
66	Mecklenburg	2018	4.67	USA	162	Chronic knee pain	A remotely delivered, home-based 12-week intervention program	Three education pieces	Knee injury and OA outcome score pain subscale	84

IA: intra-articular, NASHA: non-animal stabilized hyaluronic acid, FIHOA: Functional Index of Hand Osteoarthritis, GP: General Practitioner, EQ-5D: EuroQol five-dimension scale, SF-36: the 36-Item Short Form Health Survey, SR: slow release, DAS28-ESR: Disease Activity Score 28–joint count - erythrocyte sedimentation rate, MS: multiple sclerosis.