ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 847−855

Journal of Clinical Epidemiology

Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study

Bruno R. da Costa[a,b], Eveline Nüesch[a,b,c], Anne W. Rutjes[a,d], Bradley C. Johnston[e,f], Stephan Reichenbach[a,b], Sven Trelle[b], Gordon H. Guyatt[e], Peter Jüni[a,b,*]

[a]*Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland*
[b]*Clinical Trials Unit Bern, Bern University Hospital, Bern, Switzerland*
[c]*Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, University of London, London, UK*
[d]*Centre for Aging Sciences, G. d'Annunzio University Foundation, Chieti, Italy*
[e]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada*
[f]*Department of Anesthesia and Pain Medicine, The Hospital for Sick Children, Toronto, Canada*

Accepted 18 March 2013; Published online 6 June 2013

Abstract

Objective: To investigate whether it is valid to combine follow-up and change data when conducting meta-analyses of continuous outcomes.

Study Design and Setting: Meta-epidemiological study of randomized controlled trials in patients with osteoarthritis of the knee/hip, which assessed patient-reported pain. We calculated standardized mean differences (SMDs) based on follow-up and change data, and pooled within-trial differences in SMDs. We also derived pooled SMDs indicating the largest treatment effect within a trial (optimistic selection of SMDs) and derived pooled SMDs from the estimate indicating the smallest treatment effect within a trial (pessimistic selection of SMDs).

Results: A total of 21 meta-analyses with 189 trials with 292 randomized comparisons in 41,256 patients were included. On average, SMDs were 0.04 standard deviation units more beneficial when follow-up values were used (difference in SMDs: −0.04; 95% confidence interval: −0.13, 0.06; $P = 0.44$). In 13 meta-analyses (62%), there was a relevant difference in clinical and/or significance level between optimistic and pessimistic pooled SMDs.

Conclusion: On average, there is no relevant difference between follow-up and change data SMDs, and combining these estimates in meta-analysis is generally valid. Decision on which type of data to use when both follow-up and change data are available should be prespecified in the meta-analysis protocol.

Keywords: Meta-analysis; Change; Follow-up; Bias; Continuous outcome; Review

1. Introduction

Many trialists use continuous variables, such as pain intensity or depression severity scores, as clinical outcomes. Variables can be assessed at baseline and follow-up, and estimated treatment effects can be derived either from between-group differences in changes from baseline to follow-up or from a simple comparison of values at follow-up. Point estimates of treatment effects derived from these two approaches are identical if mean baseline values of continuous variables are the same, but will differ if there are baseline imbalances [1]. Standard deviations as measures of distribution of scores will be generally similar if the average correlation between baseline and follow-up values is approximately 0.5. If the correlation is higher than 0.5, then the standard deviations of change data will be smaller; if the correlation is lower than 0.5, then the use of change data will add variation and their standard deviation will be larger than the standard deviation of follow-up data [1]. Differences in point estimates and differences in standard deviations will both affect the estimated standardized mean difference (SMD), expressing differences in point estimates in units of the pooled standard deviation. Significance levels are derived from *t*-values, which in turn are calculated from the observed difference in point

Funding: This project is supported by a grant from the ARCO Foundation, Switzerland.

Conflict of Interest: None.

* Corresponding author. Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, Bern 3012, Switzerland. Tel.: +41 (0)31 631 57 93; fax: +41 (0)31 631 35 20.

E-mail address: juni@ispm.unibe.ch (P. Jüni).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.03.009

What is new?

- On average, there is no relevant difference between standardized mean differences (SMDs) derived from follow-up and change data.
- The Cochrane Handbook currently advises against combining SMDs derived from follow-up data and SMDs derived from change data in a single meta-analysis. The present study is the first to compare these two types of SMDs.
- Our results suggest that it is generally valid to pool SMDs derived from follow-up data and SMDs derived from change data in a single meta-analysis. This results in an increase of statistical precision of pooled estimates and allows the examination of potential sources of heterogeneity in the complete set of trials.

estimates divided by its standard error. Therefore, they will again be influenced by both baseline imbalances and correlation of baseline and follow-up data.

Meta-analysts can use either type of data (follow-up or change) to derive treatment effects from trials included in pooled analyses. In a recent analysis of 10 protocols randomly selected from the Cochrane Library, however, only four protocols specified which one would be used for the calculation of treatment effects [2]. The Cochrane Handbook does not specify which method is preferable, presenting the possibility of data-driven choice of type of data for extraction source. Furthermore, the Cochrane Handbook currently advises against combining the two types of data in a single meta-analysis. If some studies provide only follow-up data and others only change data, the Cochrane Handbook guidance prevents use of all the data. This potentially compromises statistical precision of pooled estimates and prevents the examination of potential sources of heterogeneity in the complete set of trials. If, however, SMD estimates are on average similar in follow-up and change data, power in the analysis will be gained, without introducing bias, by use of all the data.

In addition, the decision to analyze either follow-up or change data may be post hoc and data driven, but the magnitude of bias introduced in meta-analyses by the systematic extraction of data that indicates the largest treatment effect (optimistic selection of SMDs) or, conversely, the data that indicates the smallest treatment effect in each trial (pessimistic selection of SMDs) is unclear. Using data from a meta-epidemiological study of osteoarthritis trials [3–5], we therefore compared SMDs, mean differences, standard deviations, and significance levels of treatment effects of meta-analyses and their component trials derived from follow-up and change data and determined the extent of bias introduced by optimistic or pessimistic post hoc selection of either type of data in meta-analyses.

2. Methods

2.1. Selection of meta-analyses and component trials

Details of the methods used in this meta-epidemiological study are reported elsewhere [4]. We searched The Cochrane Library, Medline, Embase, and CINAHL using database-specific search strategies [4]. We included meta-analyses of randomized or quasi-randomized trials in patients with osteoarthritis of the knee or hip, which assessed patient-reported pain comparing any intervention with placebo, sham, or a nonintervention control. Reports of all component trials were obtained, without language restrictions. Two independent reviewers screened the reports for eligibility in duplicate. Disagreements were resolved by discussion.

2.2. Data extraction

The following data were extracted from trial reports, namely type of intervention, funding, publication year, publication language, design, study size, blinding of patients, losses to follow-up, exclusions, handling of missing data, and treatment effects. We approximated the means and measures of dispersion from figures whenever necessary. For crossover trials, we extracted data only from the first phase. The primary outcome was patient-reported pain associated with knee or hip osteoarthritis. If different pain assessment instruments were reported, we used a hierarchy previously described to decide which instrument to extract [6]. If more than one time point was reported, we extracted outcome data at 3 months after the end of treatment for potentially structure-modifying agents and at 12 months after the end of treatment for behavior-changing interventions. For all other interventions, we extracted outcome data at the end of the treatment. Definitions used for concealment of allocation, blinding of patients, and completeness of data analysis are provided elsewhere [4]. Two independent reviewers extracted data in duplicate. Disagreements were resolved by discussion.

2.3. Statistical analysis

We expressed the treatment effects as SMDs by dividing the between-group difference in mean values by the pooled standard deviation. The pooled standard deviation was calculated as follows:

$$\mathrm{sd}_{\mathrm{pooled}} = \sqrt{\frac{(\mathrm{n}_{\mathrm{exp}} - 1)\mathrm{sd}_{\mathrm{exp}}^2 + (\mathrm{n}_{\mathrm{con}} - 1)\mathrm{sd}_{\mathrm{con}}^2}{\mathrm{n}_{\mathrm{exp}} + \mathrm{n}_{\mathrm{con}} - 2}}$$

where sd_{exp} and sd_{con} are standard deviations in experimental and control groups, and n_{exp} and n_{con} are the number of patients analyzed. For each trial, we estimated SMDs based

on differences in changes from baseline to follow-up and SMDs based on differences in values at follow-up. Negative SMDs indicate a beneficial effect of the experimental intervention. If a trial yielded more than one randomized comparison, for example, a three-arm trial yielding one randomized comparison of celexoxib vs. control and a second randomized comparison of paracetamol vs. control [7], we inflated the standard error of mean values in the control group by the square root of the number of comparisons to account for the use of the control group in multiple comparisons. We pooled follow-up and change data SMDs across trials using inverse-variance random-effects meta-analysis, and calculated the DerSimonian and Laird estimate of the variance τ^2 as a measure of between-trial heterogeneity [8]. A τ^2 of 0.01 was considered to represent small, 0.04 moderate, and 0.12 large variability between trials.

To determine the differences in SMDs, mean differences, standard deviations, and *P*-values, we restricted the analysis to the 51 randomized comparisons (38 trials) with complete follow-up and change data available. Comparisons that required approximations to derive SMDs and trials that reported only estimates after statistical adjustments for baseline values (e.g., least-square means from analysis of covariance or estimates from linear regression model adjusted for pain intensity at baseline) were excluded. We derived pooled within-trial differences between SMDs derived from follow-up and change data for each meta-analysis, and subsequently combined the pooled difference in SMDs across meta-analyses. The SMDs from follow-up and change data originated from the same patients and were therefore correlated. Accordingly, we used a random-effects meta-regression model with robust variance estimation, which accounted for the correlation of the data within trials, to derive summary differences in SMDs as previously described [9,10]. Negative differences in SMDs indicate that follow-up data result in more beneficial SMDs than change data. The design factor defined as the standard error adjusted for the correlation within trials divided by the naïve standard error was 1.40. The τ^2 estimate of the model reflected the between-trial variation in SMDs as the measure of treatment effect rather than the between-trial variation in difference in SMDs as the parameter of interest. Therefore, we approximated τ^2 estimates for the difference in SMDs from a conventional random-effects meta-analysis of differences in SMDs after inflating the corresponding standard errors with the design factor. We then performed stratified analyses according to the prespecified characteristics of trials, namely risk of bias (blinding of patients, concealment of allocation, and analysis according to the intention to treat principle), year of publication (trials published in 1980−1998 vs. trials published in 1999−2007), sample size (small trials [$<$100 patients per group] vs. large trials [$\geq$100 patients per group]), and the type of intervention assessed in the meta-analysis (drug vs. other interventions and conventional vs. complementary medicine). All stratified analyses were accompanied by two-sided tests for interaction.

Then, we calculated the differences between follow-up and change data separately for mean differences and corresponding standard deviations. Mean differences and standard deviations of follow-up and change data were standardized in units of the pooled standard deviation of the difference in follow-up values to ensure comparability of estimates across the two types of data used: the mean differences and standard deviations were therefore divided by the pooled standard deviation of the difference in follow-up values irrespective of the type of data they originated from (follow-up or change data). To compare *P*-values, we calculated ratios, dividing *P*-values derived from follow-up data by *P*-values derived from change data, with a ratio of 1 indicating identical *P*-values. Because the play of chance is more pronounced in small than large trials, we examined the differences between follow-up and change data separately for small and large trials, again with a prespecified cutoff of 100 patients per group to classify trials according to their size. Distributions were compared between small and large trials using box and whisker plots accompanied by Levene's test modified by Brown and Forsythe [11,12] for equality of distributions around the median.

To determine the extent of bias introduced by a data-driven optimistic or pessimistic post hoc selection of either type of data in meta-analyses, we used all 292 randomized comparisons available (189 trials). If the required data were unavailable, we used approximations as previously described [13]. Whenever required, we calculated standard deviations from standard errors, confidence intervals (CIs), *P*-values, and *t*-values. If needed, we approximated the standard deviations of follow-up data from that of baseline and change data, and approximated the standard deviations of change data from that of baseline and follow-up data, assuming a correlation of 0.5 between baseline and follow-up data. For the pessimistic post hoc selection, we chose the SMD that indicated the smallest treatment benefit if both follow-up and change data were available; for the optimistic post hoc selection, we chose the SMD that indicated the largest benefit. Then, we pooled all trials available for each meta-analysis, regardless of whether the trial had allowed pessimistic and optimistic post hoc selection of SMDs. The differences between pessimistic and optimistic meta-analyses were classified based on differences in pooled estimates of treatment effects and changes in significance levels. Differences in pooled estimates were considered relevant if they were at least 0.2 standard deviation units, which corresponds to a small effect size according to Cohen [14]. Changes in significance levels were considered relevant if *P*-values crossed at least one of the four prespecified cutoffs (0.10, 0.05, 0.01, and 0.001). The analysis of bias introduced by an optimistic or pessimistic post hoc selection of data was repeated after a restriction of meta-analyses to large trials including at least 100 patients per group. All *P*-values are two sided. Analyses were performed in Stata Release 12 (StataCorp, College Station, TX).

3. Results

3.1. Characteristics of the included studies

Previous reports describe the study sample and its origin [3,4]. A total of 21 meta-analyses with 189 trials with 292 randomized comparisons in 41,256 patients were eligible. Table 1 describes the characteristics of the meta-analyses. The median number of trials per meta-analysis was seven (range: 2−29), and the median number of patients per meta-analysis was 1,430 (range: 172−14,579). The pooled treatment effect calculated from follow-up data ranged from −0.05 to −1.37 and the between-trial heterogeneity from a τ^2 of 0.00−1.87; the pooled treatment effect calculated from change data ranged from −0.03 to −0.99 and the between-trial heterogeneity from a τ^2 of 0.00−0.41. Nine meta-analyses assessed drug interventions, whereas 12 assessed nondrug interventions. Ten meta-analyses assessed interventions in complementary medicine, whereas 11 assessed interventions in conventional medicine.

For 34 (12%) randomized comparisons, approximations were required to derive the standard errors of both, differences in follow-up and change data, and for 89 (30%) and 114 (39%) randomized comparisons to derive the standard error of differences in follow-up and change data values, respectively. From the 55 (19%) remaining comparisons, which did not require approximations, 51 were included in the analysis of differences in SMDs derived from follow-up and change data, whereas 4 were excluded because they reported estimates adjusted for baseline values. Table 2 presents the characteristics of the 51 included randomized comparisons as compared with the 241 randomized comparisons that were excluded from the analysis because of approximations. Included randomized comparisons were published more recently, funded more often by nonprofit organizations, and had more complete reporting of primary outcome and sample size calculations ($P \leq 0.043$).

3.2. Comparison of SMDs derived from follow-up and change data

Figure 1 shows a Forest plot of differences in SMDs derived from follow-up and change data in the 51 randomized comparisons. The average difference was near null (difference in SMDs: −0.04; 95% CI: −0.13, 0.06; $P = 0.44$), with no evidence for variability in differences in SMDs across meta-analyses ($\tau^2 = 0.00$) or trials ($\tau^2 = 0.00$) over and above of what would be expected by chance. The corresponding median difference between follow-up and change was near zero for mean differences (median: −0.01; interquartile range [IQR]: −0.12, 0.14), standard deviations (0.00; IQR: −0.27, 0.10), and SMDs (−0.02; IQR: −0.15, 0.14), and the corresponding median ratio of *P*-values was near 1 (0.99; IQR: 0.52, 1.91). Figure 2 shows that for mean differences and SMDs, the random variation was more pronounced for small than for large trials (*P*-values for difference in variation: ≤0.011), again with differences scattered around zero; for standard deviations and *P*-values, there was little evidence for a difference in variation between small and large trials (*P*-values for difference in variation: ≥0.51). Supplementary Fig. 1 (in Appendix at www.jclinepi.com) shows that stratified analyses provided no evidence for differences in SMDs depending on the characteristics of randomized comparison.

3.3. Comparison between optimistic and pessimistic selection of SMDs

A total of 292 randomized comparisons from 21 meta-analyses were included in the analysis of optimistic vs. pessimistic SMD. For 264 randomized comparisons, SMDs could be directly derived or approximated from both follow-up and change data. Table 3 presents the results of these 21 meta-analyses after selection of the more optimistic (left) and the more pessimistic SMDs (right). There was a relevant shift in SMDs by 0.20 or higher standard deviation units in seven meta-analyses (33%). For four meta-analyses (19%), both pooled SMDs and corresponding *P*-values showed a relevant shift; for three (14%) meta-analyses, the SMD differed to a relevant extent, but significance levels remained approximately constant. For six (29%) meta-analyses, the significance level changed, but the magnitude of the SMD remained approximately constant; for the remaining 8 (38%) meta-analyses, there was no relevant shift according to our criteria. Table 4 shows the results of 13 meta-analyses that remained after a restriction to large trials only, again after selection of the more optimistic (left) and the more pessimistic SMDs (right). For none, there was a relevant shift in SMDs, whereas in eight (62%), there was a relevant shift in significance levels.

4. Discussion

In our meta-epidemiological study of osteoarthritis trials with pain scores as clinical outcomes, we found no evidence for systematic differences between estimates derived from follow-up and change data. Differences in SMDs between follow-up and change data, mean differences and standard deviations measured on original scales, and corresponding *P*-values were all scattered around the null. This suggests that there are no a priori reasons that prevent the pooling of SMDs derived from follow-up and change values in a single meta-analysis of clinical osteoarthritis trials. We believe that our results are likely to apply also to other trials using clinical scores to measure symptom severity.

In contrast, the Cochrane Handbook for Systematic Reviews of Interventions recommends against combining follow-up and change data as SMDs [15]. The rationale is that the pooled standard deviation used as denominator

Table 1. Characteristics of the included meta-analyses

Interventions	Drug intervention	Complementary medicine	Number of trials[a]	Number of patients[a]	SMD follow-up (95% CI)	Heterogeneity τ^2 (*P*-value)	SMD change (95% CI)	Heterogeneity τ^2 (*P*-value)
Acupuncture	No	Yes	6	1,725	−0.56 (−0.83, −0.28)	0.12 (≤0.001)[b]	−0.48 (−0.78, −0.18)	0.12 (≤0.001)[b]
Aquatic exercise	No	No	4	599	−0.13 (−0.29, 0.03)	0.00 (0.66)	−0.17 (−0.37, 0.03)	0.01 (0.28)
Avocado soybean	No	Yes	2	325	−0.41 (−0.63, −0.19)	0.00 (0.56)	−0.40 (−0.69, −0.11)	0.02 (0.19)
Balneotherapy	No	Yes	3	273	−1.37 (−2.72, −0.02)	1.87 (≤0.001)[b]	−0.99 (−1.62, −0.37)	0.23 (0.047)[b]
Capsaicin	No	Yes	4	284	−0.32 (−0.57, −0.07)	0.01 (0.34)	−0.42 (−0.65, −0.18)	0.00 (0.84)
Chondroitin	Yes	Yes	19	3,751	−0.76 (−0.98, −0.54)	0.23 (≤0.001)[b]	−0.77 (−1.00, −0.55)	0.24 (≤0.001)
Corticosteroids	Yes	No	3	242	−0.35 (−0.61, −0.10)	0.00 (0.39)	−0.29 (−0.74, 0.16)	0.10 (0.12)
Diacerin	Yes	No	7	1,821	−0.27 (−0.38, −0.15)	0.01 (0.27)	−0.29 (−0.45, −0.13)	0.04 (0.014)
Exercise	No	No	17	2,323	−0.30 (−0.39, −0.22)	0.00 (0.45)	−0.42 (−0.57, −0.28)	0.04 (0.019)
Glucosamine	Yes	Yes	15	1,578	−0.48 (−0.76, −0.20)	0.23 (≤0.001)[b]	−0.56 (−0.82, −0.31)	0.18 (≤0.001)[b]
Laser (LLLT)	No	Yes	7	280	−0.51 (−1.00, −0.02)	0.39 (≤0.001)[b]	−0.69 (−1.12, −0.27)	0.26 (0.005)[b]
Opioids	Yes	No	13	2,925	−0.37 (−0.44, −0.29)	0.00 (0.55)	−0.38 (−0.46, −0.29)	0.00 (0.28)
Oral NSAIDs	Yes	No	29	14,679	−0.36 (−0.42, −0.30)	0.03 (≤0.001)	−0.34 (−0.39, −0.28)	0.02 (≤0.001)
Paracetamol	Yes	No	5	1,478	−0.15 (−0.26, −0.05)	0.00 (0.62)	−0.19 (−0.37, −0.01)	0.02 (0.084)
Pulsed magnetic field	No	Yes	7	496	−0.56 (−0.91, −0.22)	0.15 (0.005)[b]	−0.48 (−0.90, −0.05)	0.26 (≤0.001)[b]
Self-management	No	No	13	4,278	−0.05 (−0.11, 0.01)	0.00 (0.56)	−0.03 (−0.11, 0.04)	0.00 (0.25)
Static magnets	No	Yes	2	172	−0.44 (−0.75, −0.14)	0.00 (0.46)	−0.20 (−0.99, 0.58)	0.41 (0.002)[b]
TENS/IF	No	Yes	7	214	−0.82 (−1.11, −0.53)	0.00 (0.95)	−0.97 (−1.27, −0.67)	0.00 (0.83)
Topical NSAIDs	Yes	No	9	1,430	−0.44 (−0.60, −0.28)	0.04 (0.026)	−0.47 (−0.69, −0.25)	0.10 (≤0.001)
Viscosupplementation	Yes	No	22	2,455	−0.30 (−0.46, −0.14)	0.10 (≤0.001)	−0.27 (−0.42, −0.13)	0.07 (≤0.001)
Weight reduction	No	No	3	222	−0.05 (−0.32, 0.21)	0.00 (0.69)	−0.23 (−0.51, 0.05)	0.00 (0.76)

Abbreviations: SMD, standardized mean difference; CI, confidence interval; LLLT, low-level laser therapy; NSAIDs, nonsteroidal anti-inflammatory drugs; TENS, transcutaneous electrical nerve stimulation; IF, interferential current stimulation.

Note: The SMDs and corresponding 95% CIs were derived from random-effects meta-analyses of all trials. Negative SMDs indicate a beneficial effect of experimental intervention.

[a] Number of trials and patients totals 189 and 41,256, as eight trials were included each in two different meta-analyses.

[b] Meta-analyses considered to have high heterogeneity between trials ($\tau^2 \geq 0.12$).

Table 2. Characteristics of randomized comparisons according to approximations used to derive the standard errors of differences in follow-up and differences in change

Characteristics	Approximations required		
	Yes ($n = 241$), n (%)	No ($n = 51$), n (%)	P-value[a]
Adequate concealment of allocation	62 (26)	16 (31)	0.41
Described as double blind	165 (69)	36 (71)	0.77
Adequate blinding of patients	100 (42)	20 (39)	0.76
Intention-to-treat analysis	48 (20)	15 (29)	0.13
Published after 1999	130 (54)	40 (78)	≤0.001
Funding by nonprofit organization	45 (19)	16 (31)	0.043
Multicenter trial	143 (59)	30 (59)	0.95
Primary outcome reported	116 (48)	39 (77)	≤0.001
Sample size calculation reported	95 (39)	33 (65)	≤0.001
Drug intervention	156 (65)	29 (57)	0.29
Complementary medicine	84 (35)	15 (29)	0.46
Large trial ($n \geq 100$ patients per group)	88 (37)	16 (31)	0.49
Primary report in English	232 (96)	51 (100)	0.16

[a] P-values based on χ^2 test. Note that the 51 comparisons that did not require approximations were used for the main analysis of differences in SMDs between follow-up and change data shown in Fig. 1.

when calculating SMDs may depend on the type of the data used. As a function of the correlation between baseline and follow-up data [16], standard deviations of change data could either be systematically larger than those of follow-up data if the correlation is low, or systematically smaller if the correlation is high. In our case of osteoarthritis trials with pain scores of a fixed range as clinical outcome, resulting in symmetrical, near-normal distributions, typically without outliers, no systematic differences were found between follow-up and change values, neither

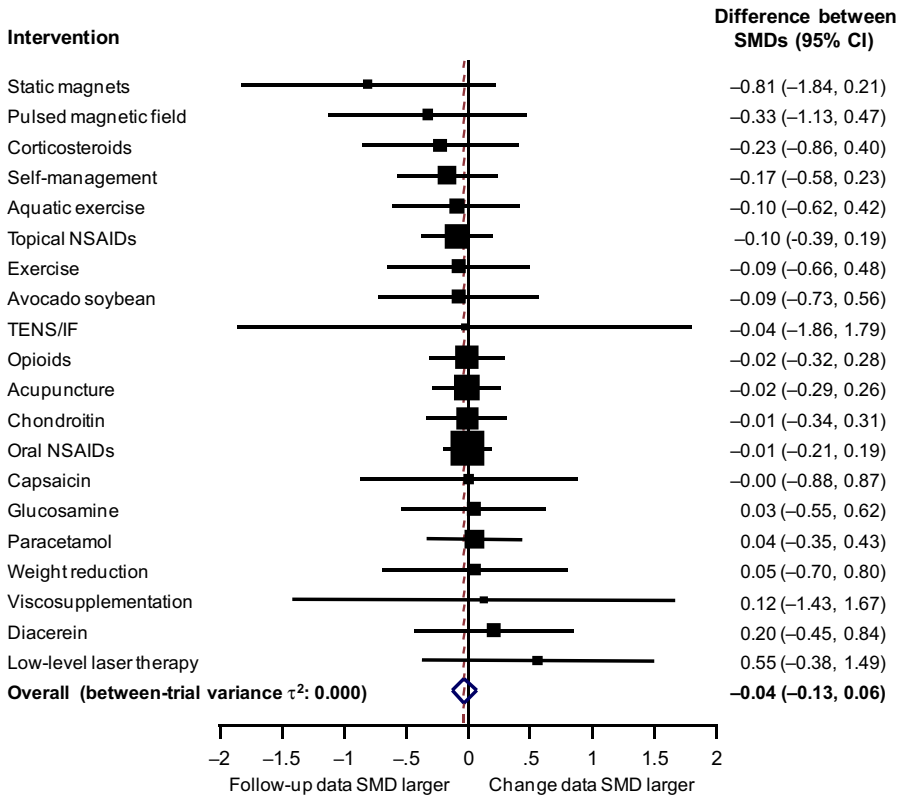

Fig. 1. Forest plot of differences in SMDs between follow-up and change data in 51 randomized comparisons included in the 20 meta-analyses. SMD, standardized mean difference; NSAIDs, nonsteroidal anti-inflammatory drugs; TENS, transcutaneous electrical nerve stimulation; IF, interferential current stimulation. Negative differences in SMDs indicate that follow-up data result in more beneficial SMDs than change data. *Note that for 7 of the 21 meta-analyses reported in Table 1, only one randomized comparison was included in the analysis of differences in SMDs between follow-up and change data, and for one intervention (balneotherapy) no trial was included.

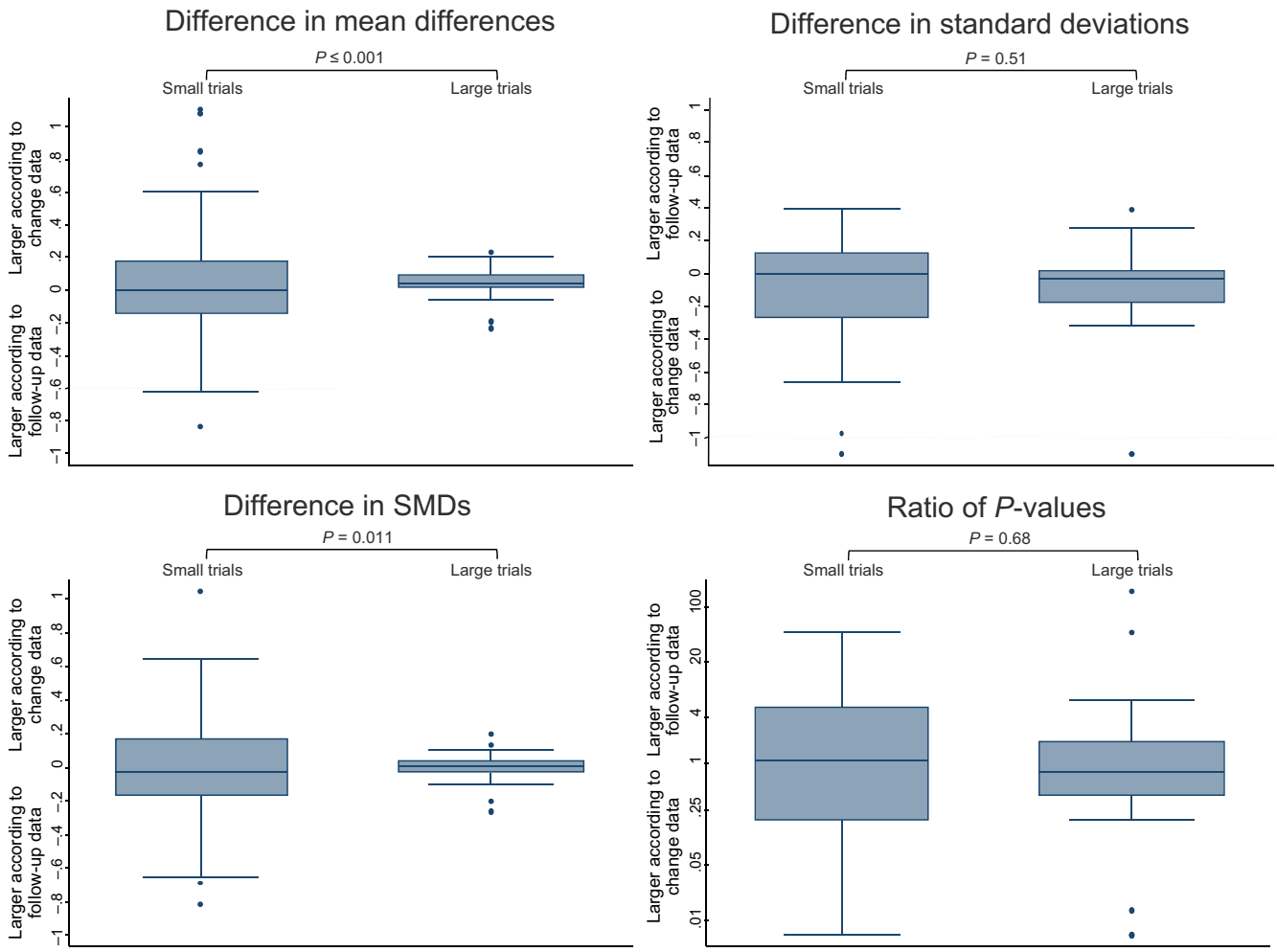

Fig. 2. Distributions of within-trial differences in SMDs, mean differences, *P*-values, and standard deviations according to sample size. SMD, standardized mean difference; small trials, <100 patients per group; large trials, ≥100 patients per group. Differences in mean differences and differences in standard deviations are expressed in units of the pooled standard deviation at follow-up.

for mean differences nor for pooled standard deviations or resulting SMDs. In the field of osteoarthritis and similar clinical settings using scores to measure symptom severity, it seems therefore justified to combine SMDs from follow-up and change data in a single meta-analysis, provided that investigators determine whether estimated treatment effects are associated with the type of data used to derive SMDs using stratified analyses accompanied by tests of interaction. In case of continuous outcomes with one-sided truncation only, such as walking distance in patients with intermittent claudication, or blood pressure in hypertensive individuals, distributions may be asymmetrical and subject to outliers, and correlations between baseline and follow-up may considerably deviate from 0.5. Future studies should investigate whether our results also apply to such outcomes.

The SMDs do not only depend on variation in the pooled standard deviations as denominator but also on random variation in the mean differences as numerator used when calculating SMDs. When comparing variation in mean differences, standard deviations, SMDs, and *P*-values between small and large trials, we found that random variation in SMDs was more pronounced in small than in large trials. This difference in variation was mainly explained by differences in mean differences, which were scattered more in small trials because of the more pronounced impact of the play of chance, which will lead to differences in baseline values of scores. This opens the possibilities for bias during data extraction in systematic reviews particularly in small trials: reviewers could consciously or unconsciously extract the more optimistic or more pessimistic values if they have personal preferences or beliefs favoring one of the compared interventions. Analyses restricted to large trials will therefore not only minimize small study effects [5] but also the impact of biased data extraction. We know of no instance, however, in which meta-analysts systematically chose, on an individual study-by-study basis, the most or least beneficial treatment effect.

For the comparison between optimistic and pessimistic selection of SMDs, we included all available 292

Table 3. Meta-analyses results according to optimistic or pessimistic post hoc selection

Intervention	Optimistic selection of SMDs		Pessimistic selection of SMDs	
	SMD (95% CI)	*P*-value	SMD (95% CI)	*P*-value
Relevant shift in pooled estimate and significance level				
Balneotherapy	−1.49 (−2.86, −0.11)	0.034	−0.94 (−1.54, −0.34)	0.002
Low-level laser therapy	−0.73 (−1.15, −0.31)	0.001	−0.47 (−0.95, 0.01)	0.053
Pulsed magnetic field	−0.63 (−1.02, −0.25)	≤0.001	−0.41 (−0.79, −0.03)	0.035
Static magnets	−0.48 (−0.78, −0.17)	0.002	−0.17 (−0.93, 0.59)	0.66
Relevant shift in pooled estimate				
TENS/IF	−1.00 (−1.30, −0.69)	≤0.001	−0.80 (−1.09, −0.51)	≤0.001
Glucosamine	−0.63 (−0.90, −0.35)	≤0.001	−0.42 (−0.67, −0.17)	≤0.001
Acupuncture	−0.60 (−0.89, −0.32)	≤0.001	−0.39 (−0.63, −0.15)	≤0.001
Relevant shift in significance level				
Capsaicin	−0.43 (−0.67, −0.19)	≤0.001	−0.31 (−0.55, −0.07)	0.013
Corticosteroids	−0.39 (−0.66, −0.12)	0.004	−0.23 (−0.63, 0.18)	0.28
Viscosupplementation	−0.34 (−0.50, −0.19)	≤0.001	−0.23 (−0.37, −0.09)	0.002
Aquatic exercise	−0.21 (−0.37, −0.05)	0.010	−0.07 (−0.23, 0.09)	0.42
Paracetamol	−0.20 (−0.31, −0.08)	≤0.001	−0.13 (−0.26, 0.00)	0.057
Self-management	−0.10 (−0.17, −0.03)	0.006	0.01 (−0.05, 0.07)	0.69
No relevant shift				
Chondroitin	−0.82 (−1.05, −0.59)	≤0.001	−0.70 (−0.91, −0.50)	≤0.001
Topical NSAIDs	−0.52 (−0.72, −0.32)	≤0.001	−0.39 (−0.57, −0.20)	≤0.001
Exercise	−0.44 (−0.56, −0.32)	≤0.001	−0.28 (−0.37, −0.20)	≤0.001
Avocado soybean	−0.44 (−0.66, −0.22)	≤0.001	−0.36 (−0.58, −0.14)	≤0.001
Opioids	−0.41 (−0.50, −0.32)	≤0.001	−0.34 (−0.42, −0.27)	≤0.001
Oral NSAIDs	−0.39 (−0.45, −0.33)	≤0.001	−0.33 (−0.38, −0.27)	≤0.001
Diacerein	−0.34 (−0.48, −0.20)	≤0.001	−0.22 (−0.34, −0.11)	≤0.001
Weight reduction	−0.19 (−0.45, 0.08)	0.166	−0.02 (−0.28, 0.24)	0.88

Abbreviations: SMD, standardized mean difference; CI, confidence interval; NSAIDs, nonsteroidal anti-inflammatory drugs; TENS, transcutaneous electrical nerve stimulation; IF, interferential current stimulation.

Note: We considered clinically relevant a difference of ≥0.2 between optimistic and pessimistic SMDs. The following thresholds in *P*-values were used to define significance level shift, namely 0.1, 0.05, 0.01, and 0.001.

randomized comparisons, irrespective of whether approximations were required to derive SMDs. This included the assumption of a correlation of 0.5 between baseline and follow-up values if measures of dispersion were unavailable for either follow-up or change data. This approach tends to bias the differences between follow-up and change data toward the null. Therefore, the observed differences between optimistic and pessimistic selection of SMDs may even increase with better quality of reporting and more complete availability of both follow-up and change data.

An incidental finding was that the variation in differences in *P*-values was equally pronounced in large and small trials. This is explained by the use of a *t*-test to derive *P*-values, which is more conservative in small than in large trials.

Table 4. Meta-analyses results according to optimistic or pessimistic post hoc selection when restricting analyses to large trials only

Intervention	Optimistic selection of SMDs		Pessimistic selection of SMDs	
	SMD (95% CI)	*P*-value	SMD (95% CI)	*P*-value
Relevant shift in significance level				
Topical NSAIDs	−0.32 (−0.54, −0.10)	0.005	−0.29 (−0.51, −0.07)	0.011
Acupuncture	−0.25 (−0.40, −0.11)	≤0.001	−0.21 (−0.38, −0.04)	0.014
Glucosamine	−0.19 (−0.35, −0.03)	0.019	−0.08 (−0.24, 0.07)	0.30
Paracetamol	−0.21 (−0.35, −0.07)	0.003	−0.15 (−0.30, 0.01)	0.062
Diacerein	−0.21 (−0.39, −0.02)	0.027	−0.07 (−0.19, 0.05)	0.256
Viscosupplementation	−0.16 (−0.27, −0.04)	0.007	−0.08 (−0.21, 0.05)	0.22
Aquatic exercise	−0.17 (−0.34, 0.00)	0.051	−0.05 (−0.22, 0.12)	0.57
Self-management	−0.09 (−0.17, −0.02)	0.016	0.02 (−0.05, 0.08)	0.63
No relevant shift				
Balneotherapy	−0.77 (−1.06, −0.47)	≤0.001	−0.62 (−0.92, −0.33)	≤0.001
Opioids	−0.36 (−0.46, −0.27)	≤0.001	−0.31 (−0.41, −0.22)	≤0.001
Oral NSAIDs	−0.37 (−0.43, −0.31)	≤0.001	−0.31 (−0.37, −0.26)	≤0.001
Chondroitin	−0.28 (−0.61, 0.06)	0.108	−0.23 (−0.52, 0.06)	0.125
Exercise	−0.25 (−0.36, −0.14)	≤0.001	−0.23 (−0.34, −0.12)	≤0.001

Abbreviations: SMD, standardized mean difference; CI, confidence interval; NSAIDs, nonsteroidal anti-inflammatory drugs.

Note: We considered clinically relevant a difference of ≥0.2 between optimistic and pessimistic SMDs. The following thresholds in *P*-values were used to define significance level shift, namely 0.1, 0.05, 0.01, and 0.001.

Differences in *t*-values between follow-up and change data of the same magnitude will impact more in large than in small trials or, conversely, smaller variations in *t*-values are required in large than in small trials to achieve a specific variation in *P*-values. Along the same lines, we found that SMDs remained largely the same in pessimistic and optimistic scenarios after restricting meta-analyses to large randomized comparisons with 200 patients or more, whereas *P*-values still differed considerably. Because meta-analyses of continuous outcomes are typically overpowered, the clinical interpretation should depend on the magnitude of the SMDs and not on *P*-values. In a recent meta-analysis of viscosupplementation, for example, we found an SMD of −0.16 (95% CI: −0.26, −0.07), which corresponds to 4 mm on a 100-mm visual analog scale, the *P*-value of which, however, was lower than 0.001 [17].

To our knowledge, this is the first study to compare SMDs derived from follow-up and change data. The comparison of SMDs was based on 51 randomized comparisons in 7,784 patients, the comparison of optimistic and pessimistic scenarios on 292 randomized comparisons in 41,256 patients from 21 meta-analyses of osteoarthritis trials. Our results suggest that it may be valid to pool SMDs derived from follow-up and change values in a single meta-analysis, provided that symptom scales of a fixed range are used and differences according to type of data used to derive SMDs are explored. However, we found relevant differences in pooled estimates of treatment effect and/or significance levels between meta-analyses after an optimistic selection as opposed to a pessimistic selection of SMDs in most of the meta-analyses. This suggests the need of a priori decisions prespecified in the protocol regarding the type of data used preferentially for the calculation of SMDs when both follow-up and change data are available.

Appendix

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jclinepi.2013.03.009.

References

[1] Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123−4.
[2] Tendal B, Higgins JP, Juni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. BMJ 2009;339:b3128.
[3] Nuesch E, Reichenbach S, Trelle S, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. Arthritis Rheum 2009;61:1633−41.
[4] Nuesch E, Trelle S, Reichenbach S, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ 2009;339:b3244.
[5] Nuesch E, Trelle S, Reichenbach S, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ 2010;341:c3515.
[6] Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006;20:721−40.
[7] Pincus T, Koch G, Lei H, et al. Patient Preference for Placebo, Acetaminophen (paracetamol) or Celecoxib Efficacy Studies (PACES): two randomised, double blind, placebo controlled, crossover clinical trials in patients with knee or hip osteoarthritis. Ann Rheum Dis 2004;63:931−9.
[8] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177−88.
[9] Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect sizes. Res Synth Methods 2010;1:39−65.
[10] da Costa BR, Rutjes AWS, Johnston BC, et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol 2012;41:1445−59.
[11] Levene H. Robust tests for equality of variances. In: Olkin I, editor. Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford: Stanford University Press; 1960.
[12] Brown MB, Forsythe AB. Robust test for the equality of variances. J Am Stat Assoc 1974;69:364−7.
[13] Reichenbach S, Sterchi R, Scherer M, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. Ann Intern Med 2007;146:580−90.
[14] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
[15] Higgins JPT, Green S (editors). Meta-analyses with continuous outcomes. In: Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available at: www.cochrane-handbook.org.
[16] Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol 1989;42:1097−105.
[17] Rutjes AW, Juni P, da Costa BR, Trelle S, Nuesch E, Reichenbach S. Viscosupplementation for osteoarthritis of the knee: a systematic review and meta-analysis. Ann Intern Med 2012;157:180−91.