Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2012;**41**:1445–1459
doi:10.1093/ije/dys124

META-ANALYSIS

Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study

Bruno R da Costa,[1] Anne WS Rutjes,[1] Bradley C Johnston,[2,3] Stephan Reichenbach,[1,4] Eveline Nüesch,[1,4,5] Thomy Tonia,[1] Armin Gemperli,[4] Gordon H Guyatt[2] and Peter Jüni[1,4]*

[1]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland, [2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada, [3]Department of Anesthesia and Pain Medicine, The Hospital for Sick Children, Toronto, Canada, [4]CTU Bern, Bern University Hospital, Switzerland and [5]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, University of London, UK

*Corresponding author. Institute of Social and Preventive Medicine (ISPM), Finkenhubelweg 11, University of Bern, 3012 Bern, Switzerland. E-mail: juni@ispm.unibe.ch

Accepted 6 July 2012

Background Clinicians find standardized mean differences (SMDs) calculated from continuous outcomes difficult to interpret. Our objective was to determine the performance of methods in converting SMDs or means to odds ratios of treatment response and numbers needed to treat (NNTs) as more intuitive measures of treatment effect.

Methods Meta-epidemiological study of large-scale trials ($\geqslant$100 patients per group) comparing active treatment with placebo, sham or non-intervention control. Trials had to use pain or global symptoms as continuous outcomes and report both the percentage of patients with treatment response and mean pain or symptom scores per group. For each trial, we calculated odds ratios of observed treatment response and NNTs and approximated these estimates from SMDs or means using all five currently available conversion methods by Hasselblad and Hedges (HH), Cox and Snell (CS), Furukawa (FU), Suissa (SU) and Kraemer and Kupfer (KK). We compared observed and approximated values within trials by deriving pooled ratios of odds ratios (RORs) and differences in NNTs. ROR <1 and positive differences in NNTs imply that approximations are more conservative than estimates calculated from observed treatment response. As measures of agreement, we calculated intraclass correlation coefficients.

Results A total of 29 trials in 13 654 patients were included. Four out of five methods were suitable (HH, CS, FU, SU), with RORs between 0.92 for SU [95% confidence interval (95% CI), 0.86–0.99] and 0.97 for HH (95% CI, 0.91–1.04) and differences in NNTs between 0.5 (95% CI, −0.1 to −1.6) and 1.3 (95% CI, 0.4–2.1). Intraclass correlation coefficients were $\geqslant$0.90 for these four methods, but $\leqslant$0.76 for the fifth method by KK (P for differences $\leqslant$0.027).

Conclusions The methods by HH, CS, FU and SU are suitable to convert summary treatment effects calculated from continuous outcomes into odds ratios of treatment response and NNTs, whereas the method by KK is unsuitable.

Keywords Continuous outcome, meta-analysis, responder analysis, response, standardized mean difference

Introduction

Systematic reviews and meta-analyses of randomized trials are often used as a basis for clinical decision making. If outcomes are measured on a continuous scale, however, meta-analysts often find trials that have used a number of different instruments to measure the same underlying construct (e.g. depression, functional capacity or pain). The generation of a pooled overall estimate requires that all treatment effects are expressed in common units. The most popular approach is the use of standardized mean differences (SMDs), also known as effect sizes. SMDs are calculated by dividing observed differences in means by the corresponding standard deviation in each trial. Resulting standardized treatment effects are expressed as standard deviation units and should ensure that effects observed in different trials can be statistically combined regardless of the type of instrument used to assess clinical outcome.

Clinicians find SMDs non-intuitive and thus difficult to interpret.[1] Instead, investigators have used responder analyses,[2,3] dichotomizing continuous data based on a pre-specified cut-off score to classify patients into treatment responders, with a reduction in symptoms, which is important to patients (e.g. $\geqslant$30% decrease from baseline), and non-responders in each group. These dichotomized data could then be compared between groups using odds ratios, risk ratios, risk differences or numbers needed to treat (NNTs) or harm, all of which are likely to enhance interpretability for the clinician. Because choosing thresholds and reporting results as proportions have been widely adopted only recently, many trials, especially those published before 2000, report only continuous data. Five methods are available to approximate measures of dichotomized treatment response from SMDs or from group-specific means and corresponding standard deviations.[4–8] To our knowledge, there are no empirical evaluations of the performance of all five methods against estimates calculated from actual treatment responses observed after dichotomization of original data in a large series of randomized trials. We therefore assembled a dataset of large trials performed in patients with osteoarthritis to determine the performance of all five methods in deriving odds ratios of treatment response and NNTs.

Methods

Literature search

We searched the Cochrane Central Register of Controlled Trials for entries from 1980 until 1 December 2010. The search strategy included text words and database-specific subject headings for knee and hip osteoarthritis (Supplementary Appendix A). One reviewer (B.d.C.) screened references for eligibility; a second reviewer (B.C.J.) screened a randomly selected sample of 20% of the references. Kappa as a measure of inter-observed agreement was 0.81.

Trial selection

We used a meta-epidemiological approach using data from trials that included patients with hip and/or knee osteoarthritis. We included placebo, sham or non-intervention control RCTs. Trials using an unpredictable allocation sequence were considered randomized; trials using potentially predictable allocation mechanisms, such as alternation or the allocation of patients according to their date of birth, were considered quasi-randomized. Trials had to report changes from baseline or final values at follow-up of pain and/or global symptoms, as well as dichotomized treatment response according to pre-determined criteria to define treatment response based on the same instrument. Studies had to include an average of at least 100 patients per group, with at least 75% of included patients diagnosed with osteoarthritis of the knee or hip. A two-arm trial with 110 patients in one arm and 95 patients in the second arm, for example, was eligible. Reports of trials were restricted to English language full-text peer-reviewed publications. The included trials were categorized according to the experimental intervention: acupuncture, viscosupplementation, food supplements, oral non-steroidal anti-inflammatory drugs (NSAIDs), topical NSAIDs, opioids, serotonin-norepinephrine reuptake inhibitors (SNRIs) and miscellaneous if only one trial examined an intervention (autologous conditioned serum, balneotherapy, ginger extract, collagen hydrolysate and paracetamol).

Data extraction and quality assessment

We extracted data from individual trials using a standardized form. Two out of three reviewers (B.d.C., B.C.J., T.T.), independently extracted the data in duplicate. Disagreements were resolved by consensus; a senior reviewer (A.W.S.R.), not otherwise involved in the data extraction process, made the final decision if reviewers failed to reach consensus. Concealment of treatment allocation was considered as adequate if investigators used central randomization, sequentially numbered, sealed, opaque envelopes or coded drug packs.[9,10] Blinding of patients was considered adequate if experimental and control interventions were described as indistinguishable or if a double

dummy technique was used.[10] Analyses were considered to follow the intention-to-treat principle if all randomized patients were reported to be included in the analysis or if the reported numbers of patients randomized and analysed were identical.[11]

Standardized mean differences

For each trial, we calculated SMDs using differences in mean change from baseline and the pooled standard deviation of mean changes. If differences in mean change from baseline were unavailable, we used differences in mean final values at follow-up and respective pooled SDs. To determine whether use of final values was likely to yield similar results to mean change, we conducted an analysis of 12 trials that provided both, changes from baseline and final values at follow-up. We determined differences in SMD between the two types of data and found SMDs much the same: difference in SMDs 0.07 [95% confidence interval (95% CI), −0.04 to 0.19]. If some of the data required were not available, we used approximations as previously described.[12] SMDs were calculated as follows:

$$SMD = \frac{mean_{exp} - mean_{con}}{sd_{pooled}} \quad (1)$$

where $mean_{exp}$ and $mean_{con}$ are mean values of the outcome in experimental and control groups, and $sd_{pooled}$is the pooled standard deviation, which was in turn calculated as follows:

$$sd_{pooled} = \sqrt{\frac{(n_{exp} - 1) * sd_{exp}^2 + (n_{con} - 1) * sd_{con}^2}{(n_{exp} + n_{con}) - 2}} \quad (2)$$

where sd_{exp} and $sd_{con}$are standard deviations in experimental and control groups, and n_{exp} and n_{con}, the number of patients analysed. This formula accounts for potential between-group imbalances in number of patients and was used to calculate all SMDs in the present investigation. The following approximation may be used when number of patients in each group is approximately the same:

$$sd_{pooled} = \sqrt{\frac{sd_{exp}^2 + sd_{con}^2}{2}}$$

Conversion methods

The following sections present the methods used to convert results of continuous outcomes into dichotomized treatment response. Throughout, we refer to 'observed' values and 'approximated' values. Observed values are based on direct dichotomization of continuous data by trialists using a pre-specified cut-off score to classify patients into treatment responders and non-responders, with numbers or percentages reported in the published article. Approximated values are those derived from differences in means between groups (typically SMD) or from group means.

We used five different methods to convert continuous outcomes into dichotomized treatment effects. The first two methods by Hasselblad and Hedges[4] and Cox and Snell[5] allow the direct conversion of SMDs into odds ratios. The third method by Furukawa[8,13] allows the conversion of SMDs into group-specific risks. The fourth method by Suissa[6] uses group means to derive group-specific risks. The fifth method by Kraemer and Kupfer[7] allows the direct conversion of SMDs into risk differences. Elaborations on these methods were recently published by Thorlund *et al.*[1] and Anzures-Cabrera *et al.*[14] Methods are summarized in the following paragraphs.

Hasselblad and Hedges' method

Following Hasselblad and Hedges' method, we multiplied the SMD and its standard error by 1.81 to calculate the log odds ratio *lnOR* and the corresponding standard error se_{lnOR}.[4,15] The method is based on the assumption that mean scores in each group follow a logistic distribution (i.e. a near normal distribution) and that variances are equal between groups.

Cox and Snell's method

Cox and Snell's method is computationally similar to Hasselblad and Hedges' method, but uses a different multiplication factor. We multiplied SMDs and their standard error by 1.65 to calculate log odds ratios and the corresponding standard errors.[5,14] The method is based on the assumption that mean scores in each group follow a normal distribution and that variances are equal between groups.

Furukawa's method

Furukawa's method requires specification of a control group risk.[8,13,16] We estimated trial-specific control group risk of treatment response $risk_{con}$ as the probability of scores of included patients to be beyond the cut-off score C used to distinguish between patients with and without treatment response in a specific trial as follows:

$$risk_{con} = \Phi\left(\frac{C - mean_{con}}{sd_{con}}\right) \quad (3)$$

where Φ is the cumulative standard normal distribution.

The SMD and the control group risk $risk_{con}$ were used to derive the experimental group risk of treatment response $risk_{exp}$

$$risk_{exp} = 1 - \Phi(SMD - \Phi^{-1}(risk_{con})) \quad (4)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution.

Then, we converted risks to odds for both groups g

$$odds_g = \frac{risk_g}{1 - risk_g} \quad (5)$$

where $risk_g$ is the group-specific risk of treatment response, and derived the log odds ratio *lnOR*.

The standard error of the log odds ratio se_{lnOR} was calculated as follows[1]

$$se_{lnOR} = \sqrt{\frac{1}{e_{exp}} + \frac{1}{e_{con}} + \frac{1}{n_{exp}} + \frac{1}{n_{con}}} \quad (6)$$

where e_{exp} and e_{con} are the estimated numbers of events in experimental and control groups derived from risks $risk_{exp}$ and $risk_{con}$ and the number of patients analysed n_{exp} and n_{con}.

Suissa's method

Suissa's method is basically equivalent to Furukawa's method.[8] However, Furukawa uses the control group–specific mean and standard deviation and the cut-off score C to derive a control group risk of treatment response and then calculates the experimental group risk of treatment response based on the calculated SMD, which in turn was derived using the pooled standard deviation. Suissa uses group-specific means and standard deviations and cut-off score C to derive the risk of treatment response in both groups rather than in the control group only.[6] For the experimental group, this is as follows:

$$risk_{exp} = \Phi\left(\frac{C - mean_{exp}}{sd_{exp}}\right) \quad (7)$$

where $risk_{exp}$is the experimental group risk of treatment success, $mean_{exp}$is the mean score for the experimental group and $sd_{exp}$is its standard deviation. See (3) for corresponding calculations for the control group risk $risk_{con}$. If $sd_{exp} = sd_{con}$, then Furukawa's and Suissa's method will yield identical results. The more discrepant sd_{exp} and sd_{con} the more results will differ between Furukawa's and Suissa's method.

In addition, the estimation of the standard error of the log odds ratio se_{lnOR}used for Furukawa's method, as specified in (6), ignores that numbers of events in experimental and control groups were only estimated and not observed.[14] Conversely, Suissa took this into account and suggested calculating the standard error of log odds ratio se_{lnOR} as follows:

$$se_{lnOR} = \sqrt{\frac{var(risk_{exp})}{(risk_{exp}(1 - risk_{exp}))^2} + \frac{var(risk_{con})}{(risk_{con}(1 - risk_{con}))^2}} \quad (8)$$

where $var(risk_{\mathrm{exp}})$ is the variance of $risk_{exp}$ and $var(risk_{\mathrm{con}})$ the variance of $risk_{con}$, with the group-specific variance estimated as follows:

$$var(risk_g) = \frac{\left(1 + \frac{1}{2}\left(\frac{C - mean_g}{sd_g}\right)^2\right)\left(exp\left(-\frac{1}{2}\left(\frac{C - mean_g}{sd_g}\right)^2\right)\right)^2}{2\pi n_g} \quad (9)$$

where n_g is the group number of participants, $mean_g$ is the group mean score and sd_g is its standard deviation, and exp is the exponential function.

Kraemer and Kupfer's method

Kraemer and Kupfer's method is based on the relationship between the risk difference RD and the area under the receiver operating characteristics curve of the probability of treatment response in the control group on the x-axis against the probability of treatment response in the experimental group on the y-axis. The more the area under the curve deviates from 0.5, which indicates no difference between groups, the more pronounced the risk difference.[7] According to this method, we used SMDs to calculate the area under the receiver operating characteristics curve AUC:

$$AUC = \Phi\left(\frac{SMD}{\sqrt{2}}\right) \quad (10)$$

and the corresponding risk difference RD is calculated as follows:

$$RD = 2 * AUC - 1 \quad (11)$$

The same approach was used to estimate upper and lower limits of the 95% CI of the risk difference directly from the 95% CI of the SMD.

Calculation of odds ratios and NNTs

The methods by Hasselblad and Hedges and Cox and Snell yielded odds ratios. To derive risk differences and corresponding NNTs, we first calculated control group risk of treatment success as shown in (3), converted the control group risk $risk_{con}$ into the control group odds $odds_{con}$ as shown in (5) and multiplied the control group odds by the odds ratio to derive the experimental group $odds_{exp}$. For both, experimental and control group, we converted odds into risks as follows:

$$risk_g = \frac{odds_g}{1 + odds_g} \quad (12)$$

and calculated corresponding risk differences RD.

Then, we calculated the number needed to treat NNT

$$NNT = \frac{1}{RD} \quad (13)$$

The methods by Furukawa and Suissa yielded group risks that were used to calculate risk differences. NNTs were calculated as in (13), and odds were derived as in (5) to calculate odds ratios. Kraemer and Kupfer's method yielded risk differences, and NNTs were calculated as in (13). Then, we calculated the control group risk of treatment response $risk_{con}$ as shown in (3) and subtracted $risk_{con}$ from the risk difference to derive the experimental group risk $risk_{exp}$. Finally, we converted risks into odds as shown in (5) and derived odds ratios.

Comparison between observed and approximated values

To compare approximated and observed odds ratios within each trial, we calculated the log ratio of odds ratios (LogROR) from the difference between the approximated and the observed log odds ratio. When exponentiated, a ratio of odds ratios (ROR) of 1 indicates no difference between approximated and observed estimates, a ROR >1 indicates that the approximated value overestimates, whereas a ROR <1 indicates that the approximated value underestimates the observed treatment effect. Observed and approximated odds ratios originated from the same data and were therefore correlated. Accordingly, we used a random-effects meta-regression model with robust variance estimation, which accounted for the correlation of data within trials to derive summary RORs[17]:

$$LogOR_{ij} = \alpha + \beta \text{ x method}_{ij} + \xi_j + \varepsilon_{ij}$$

for method i = 0,1 in trial j = 1,2,... n, with $\xi_j \sim N(0,\tau^2)$ and $\varepsilon_{ij} \sim N(0,\text{var}(LogOR_{ij}))$, with i = 0 representing the observed and i = 1 representing the approximated *LogOR*. τ^2 represents the variance between trials in observed LogORs, var($LogOR_{ij}$) represents the variance within trials, with robust variance estimation accounting for the correlation of LogORs within trials. The design factor (defined as the standard error accounting for the correlation divided by the naïve standard error) was 0.66. Because the τ^2 estimate in this model reflects the between-trial variation in observed LogORs as estimates of treatment effects, rather than the between-trial variation in the LogROR as parameter of interest, we used a conventional random-effects meta-analysis of the LogROR after correction of the corresponding standard error with the design factor and approximated τ^2 for LogRORs from the restricted maximum likelihood estimator.

To determine whether results differed according to characteristics of clinical outcomes, we performed stratified analyses according to the following pre-specified characteristics: type of instrument (visual analogue scale for pain overall, WOMAC pain subscale, patient global assessment and other instruments if used in at least two of included trials); baseline risk, i.e. the percentage of patients with treatment response in the control group (⩽20%, >20–⩽40%, >40–⩽60%, >60%); stringency of cut-off score used to define treatment response (>20–⩽40%, >40–⩽60%, >60–⩽80% or >80% change from baseline). Then, we conducted stratified analyses according to pre-specified characteristics of trials for the most cited method by Hasselblad and Hedges[4]: treatment benefit observed in the trial (small [SMD > −0.5] versus large [SMD ⩽ −0.5]); type of intervention (drug versus other interventions; complementary medicine versus other interventions); trial size (<200 patients per group versus ⩾200 patients per group); risk of bias (blinding of patient and therapist; concealment of allocation; analysis according to the intention-to-treat principle). Stratified analyses were accompanied by two-sided tests for interaction between characteristics and the logROR and tests for linear trend in case of ordered groups using random-effects meta-regression models with robust variance estimation.[17] Then, we derived summary differences in risk differences using random-effects meta-regression with robust variance estimation and the corresponding τ^2 for differences in risk differences using conventional meta-analysis as described above. The design factor was 0.62. A positive difference indicates that the approximated value overestimates the treatment effect. For both, logRORs and differences in risk differences, we calculated 95% prediction intervals (PI)[18] using the restricted maximum likelihood estimator of τ^2 for LogRORs and differences in risk differences. The 95% PI indicates the interval in which LogRORs or differences in risk differences of future trials will fall with 95% probability.

To compare NNTs, we calculated differences between approximated and observed NNTs. A positive difference indicated higher approximated NNTs than observed, hence an underestimation of the treatment effect. Differences in NNTs were not normally distributed, therefore we bootstrapped the median difference using bias correction and acceleration[19] to derive summary estimates and corresponding confidence intervals. For both odds ratios and NNTs, we graphically compared measures using scatter plots of observed versus approximated estimates with sizes of circles proportional to the inverse of the variance of observed estimates, and calculated corresponding intraclass correlation coefficients (ICCs) as measures of agreement.[20] The 95% CIs of individual ICCs and *P*-values for pairwise comparisons of ICCs were derived using bootstrap resampling.

We also approximated odds ratios and NNTs from summary SMDs observed in the seven meta-analyses of interventions with two or more trials available: oral NSAIDs, topical NSAIDs, food supplements, acupuncture, opioids, SNRI, viscosupplementation. For each of these meta-analysis, we first derived a summary SMD using a DerSimonian and Laird random-effects model[21] and then converted it into odds ratios and NNTs as described earlier in the text. To derive summary odds ratios of observed treatment response, we pooled trial-specific odds ratios for each meta-analysis using the same model. To derive summary NNTs based on observed treatment response, we first derived a summary risk ratio from trial-specific estimates for each meta-analysis. This summary risk ratio was multiplied with the median control group risk of treatment response observed in included trials to estimate the risk of treatment response in patients receiving the experimental intervention.[22] Finally, we calculated risk differences between the estimated risk of treatment response in patients receiving the

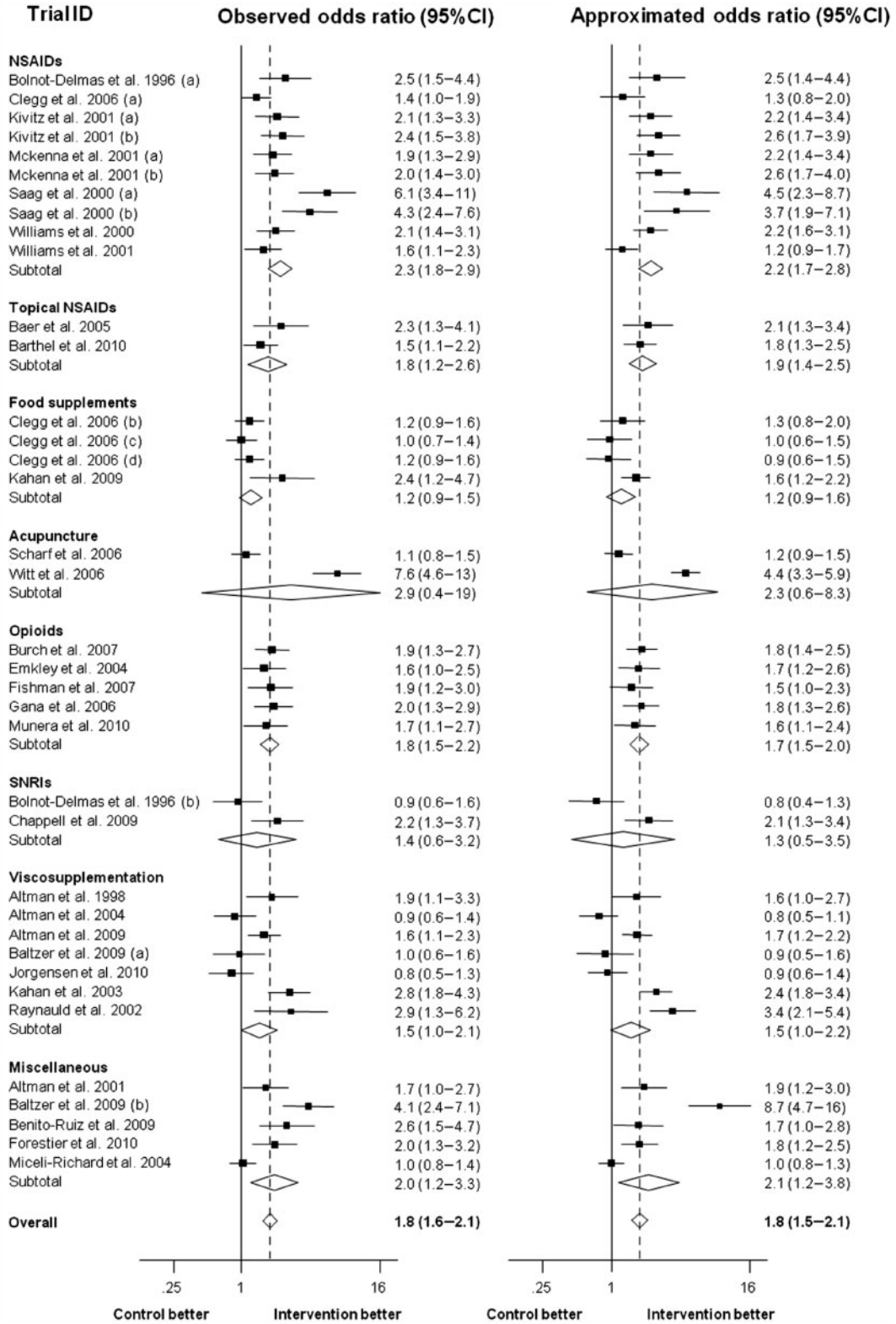

Figure 1 Forest plot showing observed odds ratios and odds ratios approximated with Hasselblad and Hedges method. Analysis is based on change from baseline values. CI, confidence interval; NSAID, non-steroidal anti-inflamatory drug; SNRI, serotonin and norepinephrine reuptake inhibitor. Designators (a) to (d) identify multiple randomized comparisons from a single trial. Note that five trials contributed with two randomized comparisons and one trial contributed with four

experimental intervention and the median control group risk and derived NNTs from the reciprocal of the risk difference. All *P*-values are two-sided. Analyses were performed in Stata Release 11 (StataCorp, College Station, TX).

RESULTS

Characteristics of included studies

Our search yielded 5290 references for screening (Supplementary Appendix B). Thirty-five reports describing 29 trials satisfied eligibility criteria. Supplementary Appendix C shows the characteristics of the included trials. In total, 13 654 patients contributed to the analysis. The treatment duration ranged from 1 day to 103 weeks (median 6 weeks), the mean age of patients from 57 to 70 years (median 62 years) and the percentage of females from 47% to 92% (median 65%). Thirteen trials (45%) reported adequate concealment of allocation, patients were appropriately blinded in 21 trials (72%) and analyses were performed according to the intention-to-treat principle in four trials (14%).

Conversion of continuous outcome into odds ratio

Figure 1 presents odds ratios of treatment response as observed (left) and as approximated from SMDs according to Hasselblad and Hedges based on differences in changes from baseline.[4] For all trials, observed and approximated odds ratios showed the same direction of treatment effect and much the same magnitude. Figure 2 shows scatter plots comparing observed odds ratios on the x-axis with approximated odds ratios on the y-axis for SMDs derived from mean changes of symptom scores from baseline for all five methods. Agreement between observed and approximated odds ratios as determined by ICC were $\geqslant 0.90$ for all methods, except for Kraemer and Kupfer's (ICC = 0.76), which was inferior to the four other methods (*P* values for pairwise differences in ICC all $\leqslant 0.027$). Supplementary Appendix D presents scatter plots and ICCs for odds ratios approximated from mean final values at follow-up.

Table 1 shows RORs pooled across all trials comparing approximated and observed estimates. Numerically, the approximation from mean changes from baseline according to Hasselblad and Hedges performed best, with an ROR of 0.97 (95% CI 0.91–1.04). The corresponding τ^2 estimate of the LogROR was 0.00, accordingly the 95% PI corresponded to the 95% CI. However, CIs between RORs according to different methods overlapped widely. Except for the ROR based on the approximation by Kraemer and Kupfer, all RORs were near 1 with a τ^2 of 0.00 and indicated that approximated odds ratios were on average somewhat more conservative than the reported data of observed treatment response. The ROR based on the approximation by Kraemer and Kupfer was 1.24 (95% CI, 1.09–1.40), reflecting an overestimation of the benefit of the experimental intervention; the corresponding τ^2 was 0.06 and the 95% PI 0.74–2.07. Supplementary Appendix E presents RORs approximated from mean final values at follow-up.

Table 2 presents stratified analyses of RORs according to probability of treatment response in the control group. For all but Kraemer and Kupfer's method, RORs were near 1 for probabilities >20–60%. For probabilities of $\leqslant 20\%$, approximated estimates became conservative, whereas for probabilities >60%, approximations became overoptimistic. However, 95% CIs overlapped widely, and tests for trend were negative. The method by Kraemer and Kupfer appeared particularly overoptimistic for probabilities of $\leqslant 40\%$, and the test for trend was positive ($P = 0.02$).

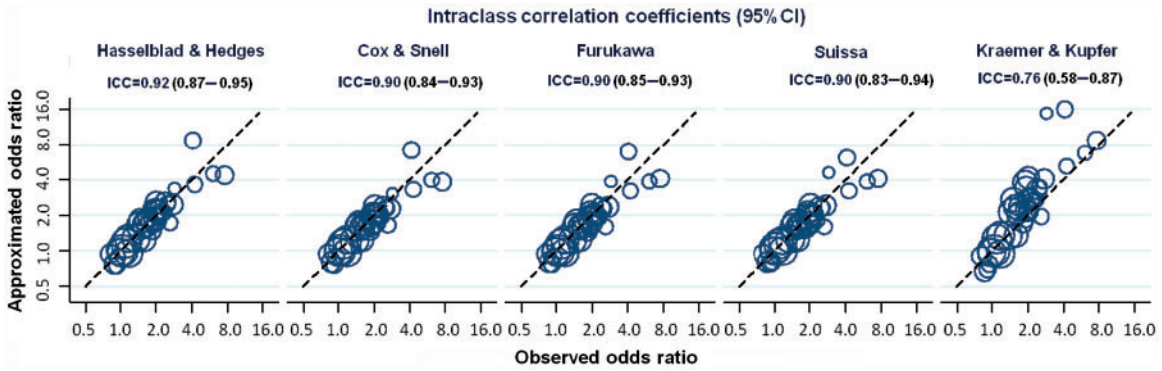

Figure 2 Scatter plots per conversion method showing the association between observed log odds ratios (x-axis) and approximated log odds ratio (y-axis) at the trial level. ICC, intraclass correlation coefficient; dashed lines indicate the line of identity between approximated and observed odds ratios; estimates lying above the line of identity indicate that the approximated odds ratio overestimates the observed treatment benefit; estimates lying below the line of identity indicate that the approximated odds ratio underestimates the observed treatment benefit. Approximated odds ratios were derived from change from baseline values; see Supplementary Appendix D for estimates based on final values at follow-up

Table 3 presents stratified analyses of RORs according to the stringency of the thresholds used to define treatment response—the extent of symptom reduction required for a patient to be considered a treatment responder. For all methods except Kraemer and Kupfer's,[7] RORs were ~1 for all thresholds. Kraemer and Kupfer's approximation became increasingly overoptimistic with more extreme thresholds used to define treatment response (test for interaction $P = 0.08$).

Table 4 presents a stratified analysis according to type of instrument used to assess symptom severity. There was some variation across instruments for all methods, but confidence intervals overlapped widely, and tests for interaction between ROR and type of instrument were negative (P for interaction $\geqslant 0.23$). For all methods, except Kraemer and Kupfer's, approximated odds ratios were more conservative or much the same as observed odds ratios, with RORs close to 1. Kraemer and Kupfer's method approximations again overestimated odds ratios. Table 5 presents stratified analyses according to characteristics of interventions and trials for Hasselblad and Hedges' method based on change from baseline data. There was no evidence to suggest that RORs differed according to any of these characteristics (P for interaction $\geqslant 0.66$).

Table 6 presents differences in approximated and observed risk differences across all trials for all methods. Again, confidence intervals overlapped widely.

Table 1 Ratio of odds ratios according to each conversion method

Method of conversion	ROR (95% CI)
Hasselblad and Hedges	0.97 (0.91–1.04)
Cox and Snell	0.92 (0.86–0.99)
Furukawa	0.93 (0.87–0.99)
Suissa	0.92 (0.86–0.99)
Kraemer and Kupfer	1.24 (1.09–1.40)

ROR, ratio of odds ratios; CI, confidence interval.

ROR of 1 means no difference between approximated and observed odds ratios; an ROR >1 means that the approximated odds ratio overestimates the observed treatment response; an ROR<1 means that the approximated odds ratio underestimates the observed treatment response.

Approximated odds ratios were derived from change from baseline values; see Supplementary Appendix E for estimates based on final values at follow-up.

Table 2 Stratified analysis comparing approximated odds ratio with observed odds ratio according to observed baseline risk

Method	Observed baseline risk	Number of comparisons	Number of patients	ROR (95% CI)	P for trend
Hasselblad and Hedges	⩽20%	6	2403	0.89 (0.73–1.09)	0.78
	>20%–⩽40%	13	4723	1.05 (0.94–1.17)	
	>40%–⩽60%	17	6193	0.94 (0.86–1.04)	
	>60%	1	335	1.15 (0.77–1.72)	
Cox and Snell	⩽20%	6	2403	0.82 (0.66–1.02)	0.38
	>20%–⩽40%	13	4723	1.00 (0.90–1.10)	
	>40%–⩽60%	17	6193	0.92 (0.84–1.00)	
	>60%	1	335	1.16 (0.79–1.71)	
Furukawa	⩽20%	6	2403	0.84 (0.69–1.03)	0.73
	>20%–⩽40%	13	4723	1.01 (0.90–1.14)	
	>40%–⩽60%	17	6193	0.91 (0.83–0.99)	
	>60%	1	335	1.16 (0.75–1.80)	
Suissa	⩽20%	6	2403	0.87 (0.69–1.10)	0.70
	>20%–⩽40%	13	4723	1.01 (0.91–1.12)	
	>40%–⩽60%	17	6193	0.92 (0.85–1.00)	
	>60%	1	335	1.16 (0.77–1.75)	
Kraemer and Kupfer	⩽20%	6	2403	1.45 (1.08–1.94)	0.02
	>20%–⩽40%	13	4723	1.41 (1.14–1.76)	
	>40%–⩽60%	17	6193	1.04 (0.95–1.14)	
	>60%	1	335	1.11 (0.71–1.73)	

ROR of 1 means no difference between approximated and observed odds ratios; ROR >1 means that the approximated odds ratio overestimates the observed treatment benefit; ROR<1 means that the approximated odds ratio underestimates the observed treatment benefit.

Approximated odds ratios were derived using change from baseline values.

Table 3 Stratified analysis comparing approximated odds ratios with observed odds ratio according to cut-off score used to generate observed odds ratio

Method	Cut-off score as percentage change from baseline	Number of comparisons	Number of patients	ROR (95% CI)	*P* for trend
Hasselblad and Hedges	>20%–⩽40%	9	3208	0.96 (0.85–1.09)	0.76
	>40%–⩽60%	17	5968	0.96 (0.86–1.08)	
	>60%–<80%	4	1541	1.00 (0.79–1.26)	
Cox and Snell	>20%–⩽40%	9	3208	0.93 (0.83–1.05)	0.96
	>40%–⩽60%	17	5968	0.91 (0.81–1.02)	
	>60%–<80%	4	1541	0.94 (0.74–1.18)	
Furukawa	>20%–⩽40%	9	3208	0.92 (0.81–1.06)	0.77
	>40%–⩽60%	17	5968	0.92 (0.83–1.02)	
	>60%–<80%	4	1541	0.97 (0.72–1.31)	
Suissa	>20%–⩽40%	9	3208	0.96 (0.84–1.08)	0.79
	>40%–⩽60%	17	5968	0.94 (0.85–1.04)	
	>60%–<80%	4	1541	1.01 (0.73–1.40)	
Kraemer and Kupfer	>20%–⩽40%	9	3208	1.10 (0.96–1.27)	0.08
	>40%–⩽60%	17	5968	1.25 (1.06–1.47)	
	>60%–<80%	4	1541	1.83 (1.01–3.31)	

ROR of 1 means no difference between approximated and observed odds ratios; ROR >1 means that the approximated odds ratio overestimates the observed treatment benefit; ROR<1 means that the approximated odds ratio underestimates the observed treatment benefit.
Approximated odds ratios were derived using change from baseline values.

Except for Kraemer and Kupfer, all differences were negative with a τ^2 of 0.00 and indicated that approximated risk differences were slightly more conservative than reported. The difference between risk differences as approximated by Kraemer and Kupfer and as observed was 4.8% (95% CI 2.3–7.3), reflecting an overestimation of the benefit of the experimental intervention; the corresponding τ^2 was 0.01, and the 95% PI −16 to 25. Supplementary Appendix F shows differences in risk differences approximated from mean final values at follow-up. Figure 3 shows scatter plots comparing corresponding NNTs as observed on the x-axis with NNTs as approximated on the y-axis, for approximations derived from mean changes for all five methods. Agreement between observed and approximated NNTs as determined by ICC were again ⩾0.90 for all methods, except for Kraemer and Kupfer's (ICC = 0.73), which was inferior to the four other methods (*P* values for pairwise differences in ICC all ⩽ 0.002). Kraemer and Kupfer's method underestimated NNTs (hence showed overoptimistic effects) in case of an observed benefit of the experimental treatment and underestimated NNHs (hence showed overly pessimistic effects) in case of observed harm of the experimental treatment. Supplementary Appendix G presents scatter plots and ICCs for NNTs approximated from mean final values at follow-up.

Table 7 shows corresponding differences in NNTs between approximated estimates and the reported data of observed treatment response. Numerically, approximations according to Hasselblad and Hedges performed best, with a difference in NNTs of 0.5 (95% CI, −0.1 to 1.6). Confidence intervals between estimates were overlapping widely, however. Again, Kraemer and Kupfer's approximation performed worst, with an overestimation of the treatment benefit, i.e. lower NNTs on average than actually observed. Supplementary Appendix H presents differences in NNTs approximated from mean final values at follow-up.

Table 8 presents pooled odds ratios (top) and NNTs (bottom) as calculated from reported data of observed treatment response and approximated from SMDs for meta-analyses on the seven interventions with at least two trials included in our study: NSAIDS (6 trials, 10 comparisons, 3127 patients), topical NSAIDs (2 trials, 2 comparisons, 708 patients), food supplement (2 trials, 4 comparisons, 1887 patients), acupuncture (2 trials, 2 comparisons, 1409 patients), opioids (5 trials, 5 comparisons, 2014 patients), SNRIs (2 trials, 2 comparisons, 475 patients), viscosupplementation (7 trials, 7 comparisons, 2640 patients). All five methods performed well, including Kraemer and Kupfer's.[7]

Discussion

In this meta-epidemiological study of 37 randomized comparisons from 29 large-scale osteoarthritis trials in 13 654 patients, we found four[4-6,13] out of five

Table 4 Stratified analysis comparing approximated odds ratio with observed odds ratio according to type of instrument

Method	Outcome measure	Number of comparisons	Number of patients	ROR (95% CI)	*P* for interaction
Hasselblad and Hedges	Pain overall VAS	9	3451	0.99 (0.87–1.13)	0.75
	Patient global assessment	7	3494	1.02 (0.90–1.16)	
	WOMAC pain	6	3348	0.91 (0.79–1.04)	
	Pain on walking VAS	3	1310	0.95 (0.75–1.21)	
	WOMAC global	2	1310	0.80 (0.45–1.43)	
	Lequesne index	2	841	1.01 (0.76–1.35)	
Cox and Snell	Pain overall VAS	9	3451	0.94 (0.83–1.07)	0.86
	Patient global assessment	7	3494	0.96 (0.85–1.08)	
	WOMAC pain	6	3348	0.89 (0.78–1.01)	
	Pain on walking VAS	3	1310	0.90 (0.72–1.14)	
	WOMAC global	2	1310	0.74 (0.37–1.49)	
	Lequesne index	2	841	0.98 (0.74–1.29)	
Furukawa	Pain overall VAS	9	3451	0.94 (0.82–1.08)	0.93
	Patient global assessment	7	3494	0.95 (0.83–1.08)	
	WOMAC pain	6	3348	0.88 (0.78–1.00)	
	Pain on walking VAS	3	1310	0.90 (0.69–1.18)	
	WOMAC global	2	1310	0.76 (0.40–1.45)	
	Lequesne index	2	841	0.98 (0.72–1.34)	
Suissa	Pain overall VAS	9	3451	0.95 (0.85–1.07)	
	Patient global assessment	7	3494	0.95 (0.84–1.07)	0.93
	WOMAC pain	6	3348	0.90 (0.81–0.99)	
	Pain on walking VAS	3	1310	0.90 (0.71–1.14)	
	WOMAC global	2	1310	0.76 (0.40–1.45)	
	Lequesne index	2	841	0.97 (0.75–1.26)	
Kraemer and Kupfer	Pain overall VAS	9	3451	1.31 (1.07–1.61)	0.23
	Patient global assessment	7	3494	1.36 (1.14–1.62)	
	WOMAC pain	6	3348	0.97 (0.86–1.10)	
	Pain on walking VAS	3	1310	1.23 (0.85–1.79)	
	WOMAC global	2	1310	1.12 (0.88–1.42)	
	Lequesne index	2	841	1.19 (0.87–1.62)	

ROR of 1 means no difference between approximated and observed odds ratios; ROR >1 means that the approximated odds ratio overestimates the observed treatment benefit; ROR<1 means that the approximated odds ratio underestimates the observed treatment benefit.
Approximated odds ratios were derived using change from baseline values.

methods suitable for responder analyses, converting differences in means of pain intensity or global symptom severity between treatment groups into odds ratios of treatment response and NNT at the level of randomized trials. When comparing estimates calculated from reported data of observed treatment response with approximated estimates, we found that approximated estimates tended to be slightly more conservative than observed estimates for all methods, except for the approach suggested by Kraemer and Kupfer[7]: approximated odds ratios were 3–8% more conservative on average for these methods[4-6,13] than odds ratios of observed treatment response. The method suggested by Kraemer and Kupfer[7] resulted in an overestimation of treatment benefits and appeared unsuitable for responder analyses.

What does this mean for a specific clinical trial? In the trial by Gana *et al.*,[23–25] for example, which shows results much in line with overall estimates, the odds ratio of treatment response comparing tramadol 200 mg daily with placebo was 2.0 (95% CI, 1.3–2.9) as calculated from reported data on treatment response, and 1.8 (95% CI, 1.3–2.6) as approximated from differences in pain intensity measured on

Table 5 Stratified analysis comparing approximated odds ratio based on Hasselblad and Hedges method to observed odds ratio

Stratified analysis	Number of comparisons	ROR (95% CI)	*P* for interaction
Overall	37	0.97 (0.91–1.04)	
Treatment effect size			0.70
Small	29	0.96 (0.91–1.01)	
Large	8	1.02 (0.77–1.35)	
Drug intervention			0.89
Yes	33	0.97 (0.92–1.02)	
No	4	1.00 (0.63–1.58)	
Complementary medicine			0.97
Yes	10	0.97 (0.79–1.19)	
No	27	0.97 (0.92–1.03)	
Concealment adequate			0.98
Yes	14	0.99 (0.86–1.14)	
Unclear	23	0.96 (0.89–1.03)	
Blinding patient and therapist adequate			0.97
Yes	8	0.97 (0.86–1.11)	
No	29	0.97 (0.90–1.05)	
ITT performed			0.69
Yes	6	1.00 (0.85–1.19)	
No	31	0.97 (0.90–1.04)	
Trial size			0.66
<200 patients per group	20	0.98 (0.88–1.10)	
⩾200 patients per group	17	0.96 (0.88–1.04)	

ITT, analysis according to the intention-to-treat principle.
Drug interventions include chondroitin, glucosamine, NSAIDs, opioids, paracetamol and viscosupplementation. Interventions in complementary medicine include acupuncture, balneotherapy, chondroitin and glucosamine.
ROR of 1 means no difference between approximated and observed odds ratios; ROR >1 means that the approximated odds ratio overestimates the observed treatment benefit; ROR<1 means that the approximated odds ratio underestimates the observed treatment benefit.
Approximated odds ratios were derived according to Hasselblad and Hedges method, using change from baseline values for the analysis.

Table 6 Difference between approximated and observed risk differences according to each conversion method

Method of conversion	DRD (95% CI)
Hasselblad and Hedges	−0.8% (−2.1 to 0.5)
Cox and Snell	−1.9% (−3.1 to −0.7)
Furukawa	−1.8% (−3.0 to −0.7)
Suissa	−1.7% (−2.8 to −0.6)
Kraemer and Kupfer	4.8% (2.3 to 7.3)

DRD, difference in risk difference.
DRD of 0 means no difference between approximated and observed risk differences; DRD >0 means that the approximated risk difference overestimates the observed treatment response; DRD <0 means that the approximated risk difference underestimates the observed treatment response.
Approximated risk differences were derived from change from baseline values; see Supplementary Appendix F for estimates based on final values at follow-up.

a 100-mm visual analogue scale according to Hasselblad and Hedges.[4] This translated into an NNT of six patients to be treated with tramadol to achieve an additional treatment response as compared with placebo when directly calculated from reported data, and an NNT of seven when approximated from differences in pain intensity, both estimates again clinically equivalent. Only for two trials, we found discrepancies that might lead to differing inferences.[26,27] Both trials evaluated unconventional interventions, one found an unusually large treatment benefit compared with placebo,[27] the other a large benefit compared with a non-intervention control.[26] When excluding these two trials from the analysis, we found τ^2 estimates to decrease by ∼40% (data available on request).

At the level of meta-analyses, random variation was even smaller and approximated odds ratios and NNT were much the same as estimates calculated from reported data of observed treatment response, irrespective of the method used. Even the method by Kraemer and Kupfer, which performed unsatisfactorily on trial level, performed reasonably well. In one meta-analysis, however, the four methods that usually performed well on trial level[4-6,13] showed discrepancies that could result in misleading impressions of the magnitude of effect. This meta-analysis addressed acupuncture and included only two trials (see Figure 1); one found a small effect as compared with a sham intervention,[28] the other an unusually large benefit compared with non-intervention control.[26]

Stratified analyses according to baseline risk of treatment response suggested that all four suitable methods[4-6,13] may be somewhat too conservative for control group response rates of ⩽20% and somewhat too optimistic for rates >60%, but rates of ⩽20% or >60% were observed in only few trials, and confidence intervals were wide and tests for interaction all negative. Similarly, in stratified analyses according to the stringency of cut-off scores to define treatment response, we did not find any evidence for differences in performance of these methods.[4-6,13] For Kraemer and Kupfer's approach we found evidence that

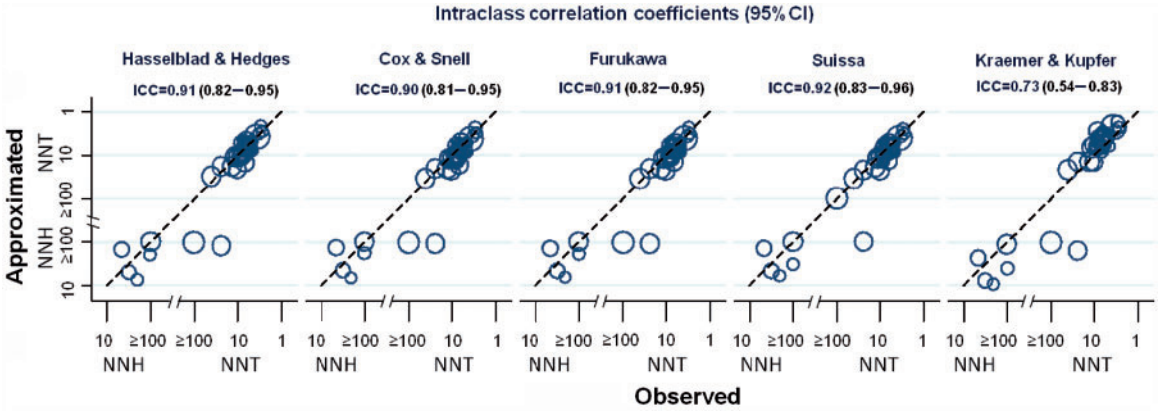

Figure 3 Scatter plots per conversion method showing the association between observed number needed to treat (x-axis) and approximated number needed to treat (y-axis) at the trial level. ICC, intraclass correlation coefficient; dashed lines indicate the line of identity between approximated and observed NNTs; estimates lying above the line of identity indicate that the approximated NNT overestimates the observed treatment benefit; estimates lying below the line of identity indicate that the approximated NNT underestimates the observed treatment benefit. Approximated NNTs were derived from change from baseline values; see Supplementary Appendix G for estimates based on final values at follow-up

Table 7 Difference between approximated and observed number needed to treat (NNT) according to each conversion method

Method of conversion	Difference in NNT (95% CI)
Hasselblad and Hedges	0.5 (−0.1 to 1.6)
Cox and Snell	1.3 (0.4 to 2.1)
Furukawa	0.9 (0.3 to 2.1)
Suissa	0.5 (0.3 to 2.2)
Kraemer and Kupfer	−1.4 (−2.2 to −1.0)

Positive differences mean that the approximated NNT underestimates the observed treatment benefit, and negative differences mean that the approximated NNT overestimates the observed treatment benefit.
Approximated NNTs were derived from change from baseline values; see Supplementary Appendix H for estimates based on final values at follow-up.

overestimations of treatment benefits increased with decreasing baseline risk of treatment response. Overestimations became particularly pronounced at baseline risks of $\leqslant$40%. As baseline risk is partially a function of the definition of treatment response, it is unsurprising that the extent of overestimation for Kraemer and Kupfer's method tended to be associated with the cut-off scores used to define treatment response.

A wide range of instruments was used to measure pain or global symptoms, and only for pain overall measured on a visual analogue scale, patient global assessment and the WOMAC pain subscale we found a sufficient number of trials to allow precise estimates; again, we did not find evidence to suggest differences in performance across instruments. Stratified analyses according to trial characteristics were performed for Hasselblad and Hedges' method only and did not suggest differences in performance of the approximations depending on these characteristics.

Our study is the most comprehensive empirical evaluation of the performance of methods used to convert continuous outcomes into odds ratios of treatment response and NNT or harm to date. As calculations of NNTs are based on risk differences, our results are also applicable to this measure of treatment benefit. The study was based on all large-scale randomized trials published as English full-text article since 1980 as identified in a systematic search of the Cochrane Central Register of Controlled Trials, which compared any intervention with placebo or non-intervention control in patients with osteoarthritis of the knee or hip and provided data on both, continuous pain or symptom severity and dichotomized treatment response. Our results may apply not only to osteoarthritis, but also to other clinical areas, particularly if scores on symptom severity are analysed, with a defined restricted range of possible scores (e.g. 0–100 mm on a visual analogue scale). This will be true if the clinical heterogeneity of patients enrolled is similar from trial to trial and not extremely homogeneous or heterogeneous, and if results approximately follow a normal distribution. Examples in which these conditions are likely to be met include depression and asthma. For outcomes that are not based on formal symptom scoring, such as blood pressure measurements in patients with arterial hypertension or walking distance in patients with intermittent claudication, the distribution of collected data is not restricted per se and skewed data could result in substantial discrepancies. Indeed, Anzures-Cabrera *et al.* found in a simulation study that most methods will result in inaccurate estimates if data are skewed or

Table 8 Odds ratio and number needed to treat according to type of intervention and conversion method based on summary SMD

Odds ratio (OR)	NSAIDs OR (95% CI)	Topical NSAIDs OR (95% CI)	Food supplement OR (95% CI)	Acupuncture OR (95% CI)	Opioids OR (95% CI)	SNRIs OR (95% CI)	Viscosupplementation OR (95% CI)
Observed treatment response	2.3 (1.8–2.9)	1.8 (1.2–2.6)	1.2 (1.0–1.4)	2.9 (0.4–19)	1.8 (1.5–2.2)	1.4 (0.6–3.2)	1.5 (1.0–2.1)
Hasselblad and Hedges	2.2 (1.7–2.8)	1.9 (1.4–2.5)	1.2 (0.9–1.6)	2.3 (0.6–8.3)	1.7 (1.5–2.0)	1.3 (0.5–3.5)	1.5 (1.0–2.2)
Cox and Snell	2.1 (1.7–2.5)	1.8 (1.4–2.3)	1.2 (0.9–1.5)	2.1 (0.6–6.9)	1.6 (1.4–1.9)	1.2 (0.5–3.2)	1.4 (1.0–2.0)
Furukawa	2.0 (1.6–2.5)	1.7 (1.4–2.2)	1.2 (0.9–1.5)	2.1 (0.6–6.8)	1.6 (1.4–1.9)	1.2 (0.5–3.1)	1.4 (1.0–2.0)
Suissa	2.1 (1.7–2.6)	1.8 (1.3–2.4)	1.2 (0.9–1.4)	2.2 (0.6–7.4)	1.7 (1.4–2.0)	1.3 (0.5–3.2)	1.5 (1.0–2.2)
Kraemer and Kupfer	2.7 (2.0–3.7)	2.2 (1.6–3.1)	1.3 (0.9–1.8)	2.8 (0.5–87)	2.0 (1.6–2.4)	1.4 (0.3–5.2)	1.6 (1.0–2.7)
Number needed to treat (NNT)	**NNT (95% CI)**	**NNT (95% CI)**	**NNT (95% CI)**	**NNT (95% CI)**	**NNT (95% CI)**	**NNT (95% CI)**	**NNT (95% CI)**
Observed treatment response	5 (4–7)	6 (3–62)	24 (12–435)	2 (0–∞)	7 (5–11)	11 (2–∞)	10 (5–2392)
Hasselblad and Hedges	5 (4–7)	6 (5–11)	21 (8–∞)	5 (2–∞)	7 (6–11)	17 (3–∞)	10 (5–∞)
Cox and Snell	6 (4–8)	7 (5–13)	23 (9–∞)	5 (2–∞)	8 (6–12)	19 (4–∞)	12 (6–∞)
Furukawa	6 (4–8)	7 (5–13)	24 (10–∞)	6 (2–∞)	8 (6–12)	19 (4–∞)	12 (6–∞)
Suissa	5 (3–7)	5 (3–11)	26 (10–∞)	3 (1–∞)	8 (5–15)	16 (3–∞)	10 (4–∞)
Kraemer and Kupfer	4 (3–6)	5 (4–9)	16 (7–∞)	4 (2–∞)	6 (5–9)	13 (3–∞)	8 (4–∞)

Analysis is based on change from baseline values.
CI, confidence interval; NSAID, nonsteroidal antiinflamatory drug; SNRI, serotonin and norepinephrine reuptake inhibitor.

standard deviations differ substantially across treatment groups.[14] To minimize the influence of small study effects due to selective reporting and publication and low methodological quality of small trials, we restricted our sample to trials that enrolled 100 patients per group.[29] Our results may, therefore, not apply for single small-scale trials, as simulations from Anzures-Cabrera *et al*. suggest.[14] When conversion methods are used in a meta-analysis of multiple trials, with an accumulated number of patients of several hundreds to a few thousands,[11] this limitation will not apply.

In 2005, Furukawa *et al.*[16] determined the performance of their own approximation method using data from 4 meta-analyses of 47 trials in 4540 patients with depression or panic disorder. Approximated risk ratios of treatment response were much the same as estimates calculated from observed treatment response, albeit slightly more conservative, as observed in our study. Furukawa and Leucht[8] subsequently determined the performance of their own method as compared with Kraemer and Kupfer's in approximating NNTs in four meta-analyses, including 10 trials of second-generation anti-psychotics in 4278 patients with schizophrenia. Consistent with our results, they found Furukawa's method more accurate than Kraemer and Kupfer's. If definitions of treatment response required a change in symptom severity of $<80\%$, Furukawa's approximation yielded NNTs that were only slightly more conservative than observed. For more stringent definitions of treatment response, with required changes from baseline of $\geqslant 80\%$, Furukawa's method became unacceptably conservative. Kraemer and Kupfer's approximation was always overoptimistic and deviated more from observed estimates with more extreme definitions of treatment response, as was the case in our study (Table 3). Comparisons of statistical methods typically involve three steps: statistical theory, simulation studies and empirical evaluations in real-world datasets. Anzures-Cabrera *et al.*[14] compared the methods by Hasselblad and Hedges,[4] Cox and Snell[5] and Suissa[6] based on statistical theory and a comprehensive simulation study, and empirically determined their performance in a convenience sample of 16 trials with 2247 patients with Alzheimer dementia or anxiety disorders. As in our study, approximated odds ratios were similar and slightly more conservative than the odds ratio of observed treatment success, with Hasselblad and Hedges' approximation being closest to the observed estimate. We believe that our study complements and extends on these studies. It compares all five methods available to date, empirically evaluates these methods in a larger dataset of 29 trials with 13 654 patients, is based on a systematic search of the literature and covers a different clinical condition.

Recent guidelines on assessment of chronic pain[30,31] and osteoarthritis,[32] as well as the US Food and Drug Administration,[33] suggest the use of responder analyses to facilitate interpretability of treatment effects measured on a continuous scale. For the purpose of this report, we presented the performance of currently available methods to approximate comparisons of responders between groups on odds ratio, NNT and risk difference scales. However, all four methods that performed well on these scales perform equally well on a risk ratio scale (data available on request). As Hasselblad and Hedges' and Cox and Snell's conversion methods directly yield odds ratios, whereas Furukawa's and Suissa's approaches yield group specific risks, the investigators' preferred scale to express treatment effects may guide the selection of conversion method.

Supplementary Data

Supplementary Data are available at *IJE* online.

Funding

This project was funded by the ARCO Foundation, Switzerland.

Acknowledgements

We thank Marcel Zwahlen and Thomas Gsponer for helpful comments and Shelagh Redmond and Pippa Scott for support in database development.

Conflict of interest: None declared.

KEY MESSAGES

- Clinicians find standardized mean differences calculated from continuous outcomes difficult to interpret.
- Standardized mean differences and means can be converted into odds ratios of treatment response and numbers needed to treat as more intuitive measures of treatment effect.
- Currently the methods described by Hasselblad and Hedges, Cox and Snell, Furukawa and Suissa are suitable to convert summary treatment effects calculated from continuous outcomes into odds ratios of treatment response and numbers needed to treat.

References

[1] Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. *Res Synth Method* 2011; **2:**188–203.

[2] Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;**316:**690–93.

[3] Dionne RA, Bartoshuk L, Mogil J, Witter J. Individual responder analyses for pain: does one pain scale fit all? *Trends Pharmacol Sci* 2005;**26:**125–30.

[4] Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;**117:**167–78.

[5] Cox DR, Snell EJ. *Analysis of Binary Data*. London: Chapman and Hall, 1989.

[6] Suissa S. Binary methods for continuous outcomes: a parametric alternative. *J Clin Epidemiol* 1991;**44:**241–48.

[7] Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry* 2006;**59:**990–96.

[8] Furukawa TA, Leucht S. How to obtain NNT from Cohen's d: comparison of two methods. *PLoS One* 2011; **6:**e19070.

[9] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;**323:**42–46.

[10] 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.

[11] 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.

[12] 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.

[13] Furukawa TA. From effect size into number needed to treat. *Lancet* 1999;**353:**1680.

[14] Anzures-Cabrera J, Sarpatwari A, Higgins JP. Expressing findings from meta-analyses of continuous outcomes in terms of risks. *Stat Med* 2011;**2011:**4298.

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

[16] Furukawa TA, Cipriani A, Barbui C, Brambilla P, Watanabe N. Imputing response rates from means and standard deviations in meta-analyses. *Int Clin Psychopharmacol* 2005;**20:**49–52.

[17] Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect sizes. *Res Synth Method* 2010;**1:**39–65.

[18] Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;**172:**137–59.

[19] Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987;**82:**171–85.

[20] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86:**420–28.

[21] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7:**177–88.

[22] Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. *BMJ* 1999;**318:**1548–51.

[23] Gana TJ, Pascual ML, Fleming RR *et al*. Extended-release tramadol in the treatment of osteoarthritis: a multicenter, randomized, double-blind, placebo-controlled clinical trial. *Curr Med Res Opin* 2006;**22:**1391–401.

[24] Florete OG, Xiang J, Vorsanger GJ. Effects of extended-release tramadol on pain-related sleep parameters in patients with osteoarthritis. *Expert Opin Pharmacother* 2008;**9:**1817–27.

[25] Kosinski M, Janagap C, Gajria K, Schein J, Freedman J. Pain relief and pain-related sleep disturbance with extended-release tramadol in patients with osteoarthritis. *Curr Med Res Opin* 2007;**23:**1615–26.

[26] Witt CM, Jena S, Brinkhaus B, Liecker B, Wegscheider K, Willich SN. Acupuncture in patients with osteoarthritis of the knee or hip: a randomized, controlled trial with an additional nonrandomized arm. *Arthritis Rheum* 2006;**54:** 3485–93.

[27] Baltzer AW, Moser C, Jansen SA, Krauspe R. Autologous conditioned serum (Orthokine) is an effective treatment for knee osteoarthritis. *Osteoarthr Cartil* 2009;**17:**152–60.

[28] Scharf HP, Mansmann U, Streitberger K *et al*. Acupuncture and knee osteoarthritis: a three-armed randomized trial. *Ann Intern Med* 2006;**145:**12–20.

[29] Nuesch E, Trelle S, Reichenbach S *et al*. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ* 2010;**341:**c3515.

[30] Dworkin RH, Turk DC, Wyrwich KW *et al*. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008;**9:**105–21.

[31] Dworkin RH, Turk DC, McDermott MP *et al*. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. *Pain* 2009;**146:**238–44.

[32] Pham T, van der Heijde D, Altman RD *et al*. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. *Osteoarthr Cartil* 2004;**12:** 389–99.

[33] U.S. Department of Health and Human Services. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes* 2006;**4:**79.