International Journal of Methods in Psychiatric Research
*Int. J. Methods Psychiatr. Res.* 22(1): 1–15 (2013)
Published online 12 March 2013 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1377

# Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation

LEVENTE KRISTON

Department of Medical Psychology, University Medical Centre Hamburg-Eppendorf, Martinistr. 52, D-20246 Hamburg, Germany

**Key words**
clinical heterogeneity, meta-analysis

**Correspondence**
Levente Kriston, Department of Medical Psychology, University Medical Centre Hamburg-Eppendorf, Martinistr. 52, D-20246 Hamburg, Germany.
Telephone +49 40 7410 56849
Fax +49 40 7410 54965
Email: l.kriston@uke.de

**Abstract**

*Objective:* There is an ongoing debate how to interpret findings of meta-analyses when substantial clinical heterogeneity is present among included trials. The aim of the present study was to demonstrate various ways of dealing with clinical heterogeneity along with underlying assumptions and interpretation. A recent meta-analysis on long-term psychodynamic psychotherapy (LTPP) was used as an illustrative example.
*Method:* Re-analysis of published data including calculation of a prediction interval, heterogeneity tests, Bayesian meta-analysis, meta-regression, and subgroup analysis to explore and interpret summary estimates in clinically heterogeneous studies.
*Results:* Meta-analytic results and their implications varied considerably depending on whether and how clinical heterogeneity was addressed.
*Conclusions:* Whether or not to trust summary estimates in meta-analysis depends largely on the subjective relevance of clinical heterogeneity present. No single analysis and interpretation strategy can be valid in every context or paradigm, thus, reflection of own beliefs on the role of heterogeneity is needed. *Copyright © 2013 John Wiley & Sons, Ltd.*

## Introduction

Evidence on effectiveness of long-term psychodynamic psychotherapy (LTPP) from randomized-controlled trials is scarce (Fonagy, 2010a). Therefore, it is of great importance that recently Leichsenring and Rabung (2011) performed a meta-analysis in order to summarize randomized evidence on effectiveness of LTPP. They included randomized-controlled trials that compared LTPP of a minimum duration of one year or 50 sessions with other psychotherapeutic treatments as active comparators. In their analysis of 10 trials in 971 patients with personality disorders, chronic disorders, or multiple mental disorders Leichsenring and Rabung (2011) found a statistically significant standardized between-group effect size (Hedges' *d*) of 0.54 for the overall effectiveness outcome and concluded that LTPP was superior to less intensive methods of psychotherapy in complex mental disorders.

Although they clearly defined the treatment of interest (LTPP), Leichsenring and Rabung (2011) adopted broad inclusion criteria on the investigated patients and comparator treatments. They defined their target population as patients with "complex mental disorders", and the investigated studies included patients with personality disorders, eating disorders, as well as depressive and anxiety disorders. Under the term "less intensive methods of psychotherapy", which was used to describe comparator treatments, cognitive (behavioural) therapy, dialectical

behavioural therapy, structured clinical management, and treatment as usual were subsumed, among others. In summary, some may consider the analysed patient populations and comparator treatments *clinically heterogeneous*. This raises important questions: How should we interpret meta-analytical findings in clinically heterogeneous trials? How much clinical heterogeneity is *too* much? Which options do we have to deal with this heterogeneity? In the case of LTPP, where a practical interpretation of the authors is largely missing: Have Leichsenring and Rabung (2011) evidenced superiority of LTPP over *all* forms of less intensive psychotherapy in *all* (or at least in the investigated) complex mental disorders? If yes, how can we use this finding to inform clinical practice? If not, what are the implications of their findings?

Although questions regarding clinical heterogeneity in meta-analysis are not new (Greenland, 1994; Thompson, 1994; Smith *et al.*, 1997; Lau *et al.*, 1998; Higgins *et al.*, 2002), a consensual answer is lacking. One of the major dichotomies concerning this topic is that of "lumping" versus "splitting". "Lumpers" tend to define their questions broadly and are rather liberal with regard to including clinically heterogeneous trials, for which they provide overall summary estimates (Gøtzsche, 2000). On the contrary, "splitters" prefer formulating narrow research questions and focus their analyses on clinically homogeneous groups of trials (Counsell, 1997). The Cochrane Handbook for Systematic Reviews of Interventions lists several advantages and disadvantages to both broad and narrow research questions (Higgins and Green, 2011). In the present study the terms "lumping" and "splitting" are used in a descriptive sense; they simply inform on how a meta-analysis was conducted.

Starting at the lumping versus splitting dichotomy the present study aims at presenting various approaches to dealing with clinical heterogeneity in meta-analysis. In addition to a pure *description* (how clinical heterogeneity can be handled), the role of underlying *assumptions* and *beliefs* behind a meta-analysis (how clinical heterogeneity is defined and perceived) and implications for *interpretation* (what the findings mean) are explored. The earlier mentioned meta-analysis on LTPP by Leichsenring and Rabung (2011) is used as an illustrative example including partial re-analysis of data.

## Lumping as literature synthesis

The rationale behind a meta-analysis is not always clearly stated. Probably, the most frequently reported motive is "to summarize evidence" on a particular topic. This may be an appropriate proximal reason for a meta-analysis, but seldom informs on why a summary is needed. This form of lumping is sometimes described as "literature synthesis" or "summary" approach (Wachter *et al.*, 1990; Rubin, 1992) and "possibly reflects an emphasis throughout medical statistics on inference rather than on decision making" (Ades *et al.*, 2005, p. 652).

In this context the choice of a statistical model to summarize effects across trials is of central interest. Usually, fixed-effect or random-effects models are considered to obtain a summary estimate with a corresponding 95% confidence interval. A *fixed-effect* model assumes that all trials estimate the same (fixed) true treatment effect, while *random-effects* models assume that no single true treatment effect exists, but each trial estimates a trial-specific true treatment effect, with these effects following a certain (commonly normal) distribution (Lau *et al.*, 1998; Higgins *et al.*, 2009; Riley *et al.*, 2011). As the Cochrane Handbook states, the random-effects model "represents our lack of knowledge about why real, or apparent, intervention effects differ by considering the differences as if they were random" (Higgins and Green, 2011, section 9.5.4). Based on the presented underlying assumptions summary estimates obtained by fixed-effect and random-effects models are conceptually different. A fixed-effect summary estimate with its 95% confidence interval describes the *true* (common) treatment effect along with its precision. A random-effects summary however informs about the *average* treatment effect and its precision.

In psychotherapy research the literature synthesis approach has a long tradition. Several authors were interested in the overall effectiveness of psychotherapeutic treatments and contributed to the development of the methodological-statistical foundations of meta-analysis (Smith *et al.*, 1980; Lipsey and Wilson, 1993). The meta-analysis on LTPP by Leichsenring and Rabung (2011) may be considered as part of this tradition of lumping psychotherapy research literature synthesis. They declared their study to be an update of a previous meta-analysis (Leichsenring and Rabung, 2008). That first analysis received several critical remarks (Thombs *et al.*, 2009; Bhar *et al.*, 2010; Littell and Shlonsky, 2010) and the authors attempted to improve their approach and provided a substantial update.

Leichsenring and Rabung (2011) used a random-effects model and calculated an effect size (Hedges' *d*) for the main outcome ("overall effectiveness", i.e. the average effect size on psychiatric symptoms, personality functioning, and social functioning) of 0.54 with a 95% confidence interval of 0.26 to 0.83. This is considered a medium-sized effect (Cohen, 1988) with a fairly convincing precision, which several researchers are likely to translate as clinically


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

relevant. Accordingly, Leichsenring and Rabung (2011) conclude that LTPP is superior to less intensive forms of psychotherapeutic interventions. Using a more exact wording one could put the results of the meta-analysis as "LTPP has a statistically significant and medium-sized *average* effect over less intensive psychotherapeutic treatments". Note that the emphasis on the *average* effect is due to the random-effects model that was implemented. The findings indicate not less and not more than that the *mean* of the distribution of the trial-specific true treatment effects is statistically different from zero.

## Lumping in a decision-making context

Imagine that you are a leading clinician treating mainly chronically depressed patients with short- to medium-term behavioural therapy and consider LTPP as a new outpatient treatment to be implemented in your department. Would you rely on the earlier described meta-analysis by Leichsenring and Rabung (2011) including various diagnostic groups and control interventions as strong evidence that LTPP will lead to improved outcomes in *your* practice? Even if you considered the differences in treatment effect estimates across the patient populations and comparator treatments investigated in the meta-analysis as *random* (they all belong to the broader classes of "complex mental disorders" and "less intensive psychotherapeutic treatments", respectively), you would presumably still want to know how large these differences are, i.e. whether and to which extent the effect of LTPP *varied* across trials.

In meta-analysis variation in the effect estimates from a set of studies is termed (*statistical*) *heterogeneity*. Whether heterogeneity is present is commonly tested with Cochran's $Q$ test statistic that examines the null hypothesis that all studies estimate the same effect (Cochran, 1954). A statistically significant result of Cochran's test (at $P < 0.05$ or with more caution at $P < 0.10$) suggests that heterogeneity is present (Higgins and Green, 2011). A frequently used descriptive measure of the percentage of statistical heterogeneity beyond the amount expected solely due to chance (sampling error) is the $I^2$ statistic that is calculated from Cochran's $Q$ and the corresponding degrees of freedom (Higgins *et al.*, 2003). According to the Cochrane Handbook an $I^2$ of 0% to 40% might not be important, 30% to 60% may represent moderate heterogeneity, 50% to 90% may represent substantial heterogeneity, and 75% to 100% may indicate considerable heterogeneity (Higgins and Green, 2011). It should be noted that the intervals are overlapping because various factors can influence the importance of the observed $I^2$ value, e.g. magnitude and direction of trial effects.

As clinical heterogeneity is a major cause of statistical heterogeneity, it is no surprise that in the meta-analysis by Leichsenring and Rabung (2011) statistical heterogeneity proved to be not only statistically significant ($Q = 24.74$, $\mathrm{df} = 9$, $P = 0.003$) but also substantial as signalized by an $I^2$ index of 64%. This indicates that two-thirds of the variability in effect estimates are owing to heterogeneity between studies rather than sampling error. These parameters were recalculated from the effect sizes of the included studies reported by Leichsenring and Rabung (2011; see also the Appendix for details). Unfortunately, in their publication Leichsenring and Rabung (2011) reported an atypical $Q$ statistic (11.72), which led to an erroneous $I^2$ estimate (23%). They are likely to have overlooked that the $Q$ value reported by the software they used (MetaWin) in case of a random-effects analysis is *not* Cochran's standard $Q$ statistic (which is calculated from fixed-effect weights and serves as basis for the calculation of the standard $I^2$), but a "generalized" variant of it based on random-effects weights and used mostly for estimating the between-trial variance component (DerSimonian and Kacker, 2007; see also the Appendix for details). Considering that not only the highly accessed and cited first meta-analysis on LTPP by Leichsenring and Rabung (2008) used MetaWin (and may have repeated the same error) but also several other researchers work with this software, special care and consulting the user's manual is advised when it comes to testing of heterogeneity with MetaWin in order to keep estimates comparable to other statistical packages.

Turning back to our decision-making example, the substantial statistical heterogeneity identified would probably make you sceptical about the expected effect of LTPP when introduced in your practice. However, you may raise the question whether the identified heterogeneity is large enough to nullify the medium-sized average treatment effect. You may still argue that although heterogeneity is disturbing, it does not really matter whether the expected effect in your department is small, moderate, or large, as long as it exists (i.e. LTPP is superior to your current practice). In order to aid decision-makers in similar situations, Riley *et al.* (2011) presented the concept of *prediction intervals*. A prediction interval describes *the expected effect of a treatment when it is applied within an individual setting* (of course, only as long as all other aspects of the individual setting are comparable to that of the analysed trials). Correspondingly, a 95% prediction interval provides the bounds for the expected treatment effect in 95% of the individual study settings. A prediction interval can easily be calculated from random-

effects models using the summary treatment effect, the standard error of the summary treatment effect, and the between-trial variance (Riley *et al.*, 2011). An assumption behind the calculation of prediction intervals is that trials are considered more or less homogeneous entities (see explanation of random-effects models earlier). In case of the LTPP meta-analysis it would mean that included patient populations and comparator treatments should be considered *exchangeable*, which is indicated if one is ready to accept "complex mental disorders" and "less intensive forms of psychotherapeutic interventions" as sufficiently homogeneous classes with regard to the effectiveness of LTPP. If it is not the case, prediction intervals will be seen just as useless as average effect estimates.

In the meta-analysis of Leichsenring and Rabung (2011) a 95% prediction interval can be calculated as −0.22 to 1.30 (for calculation see the Appendix). This interval contains values below zero, therefore, although *on average* LTPP seems effective, it may not always be beneficial *in any individual setting*. This finding suggests that the *expected* treatment effect in case of an attempted implementation of LTPP is by far not as precise as the *summary* treatment effect. In general, in a clinical decision-making context the prediction interval should be considered as additional information, because reliance solely on the summary effect size and its confidence interval could be misleading.

## Bayesian meta-analysis and the role of prior beliefs

Interpretation of any scientific (not only meta-analytic) finding is always made on the basis of existing knowledge and expectations. When a study result is judged, some kind of cognitive integration of these *prior beliefs* and the empirical data is needed. Usually, more and stronger empirical evidence is necessary to convince a sceptic than someone who was already confident that the tested hypothesis was true, and sometimes only a very large amount of unequivocal evidence can persuade an opponent.

Prior beliefs in traditional ("frequentist") meta-analysis are mainly reflected in the choice between a fixed-effect and a random-effects model. If the included trials are thought to estimate a common treatment effect, a fixed-effect model is preferred, but if heterogeneity is expected, a random-effects model can be used. Thus, the choice between these two models concerns rather the between-trial variance than the estimate of the treatment effect itself. *Bayesian meta-analysis* (Smith *et al.*, 1995; Sutton and Abrams, 2001) offers a more flexible approach, in which *prior beliefs* on the treatment effect in addition to the between-trial variance can be mathematically modelled. Consequently, in addition to the presumption in traditional meta-analysis whether between-trial variance *is present or not*, Bayesian meta-analysis allows a more exact specification of the prior *probability distribution* (e.g. mean and precision in case of a normal distribution) of the modelled parameters. Additionally, in Bayesian analysis probability statements can be made directly regarding quantities of interest, for example, the probability that patients receiving treatment A will have a better outcome than patients receiving treatment B.

Results of a Bayesian re-analysis of the LTPP data published by Leichsenring and Rabung (2011) according to various definitions of prior parameter probability distributions are presented in Table 1. All models were computed with Monte Carlo Markov chain simulation with three independent chains in WinBUGS version 1.4.3 (Lunn *et al.*, 2000). In all cases a burn-in of 20,000 simulations was discarded and presented results were obtained by a further sample of 80,000 simulations. Several prior distributions were defined for the treatment effect (including its precision) and the between-trial standard deviation parameter. Vague (uninformative) parameters mean that the defined distributions allow for a very large range of values and no substantial prior information is present. A large effect was defined as $d = 0.8$. High precision was defined as a standard error of the average treatment effect of 0.10 and indicates a high degree of certainty about the treatment effect. Low precision was defined as a standard error of 0.32. An informative between-trial standard deviation was modelled with a gamma distribution $\Gamma(1, 0.1)$ meaning that a value between 0.16 and 1.99 is expected with a 95% probability and a median at 0.38. In some analyses between-trial standard deviation was fixed at zero leading to a fixed-effect model. Some of the prior distributions were defined somewhat arbitrary but represent reasonable scenarios and provide a solid basis for judging Bayesian results on LTPP (for more details and WinBUGS code see the Appendix).

Depending on the prior assumptions on the size and precision of the effect as well as on the within-trial standard deviation (first three columns in Table 1), posterior estimates show considerable variation. Using uninformative priors (first line in Table 1), i.e. letting the data "speak", leads to estimates largely comparable with that from Leichsenring and Rabung (2011). The average treatment effect is estimated 0.54 with a 95% credible interval (corresponds to confidence interval) of 0.28 to 0.84 and a between-trial standard deviation of 0.31. The treatment effect in a new trial or individual setting is expected to fall between −0.25 and 1.36 with a probability of 95% (corresponds to the prediction interval). The probability that LTPP will perform at least as well or better than the comparator treatment is 93.5%, while an at least


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

**Table 1.** Bayesian meta-analysis with various prior assumptions

| Priors | | | Posterior estimates | | | | | |
|---|---|---|---|---|---|---|---|---|
| *d* | *Prec(d)* | *sd* | *d* | *CrI (d)* | *sd* | *CrI (d.new)* | *p1* (%) | *p2* (%) |
| vague | vague | vague | 0.54 | (0.28, 0.84) | 0.31 | (−0.25, 1.36) | 93.5 | 86.1 |
| vague | vague | none | 0.48 | (0.34, 0.61) | — | (0.34, 0.61) | 100 | 100 |
| null | low | vague | 0.46 | (0.19, 0.69) | 0.31 | (−0.37, 1.22) | 90.7 | 81.6 |
| null | low | none | 0.46 | (0.33, 0.59) | — | (0.33, 0.59) | 100 | 100 |
| null | low | informative | 0.46 | (0.19, 0.70) | 0.35 | (−0.36, 1.26) | 89.0 | 77.4 |
| null | high | vague | 0.13 | (−0.08, 0.33) | 0.58 | (−1.22, 1.36) | 61.7 | 47.0 |
| null | high | none | 0.33 | (0.22, 0.44) | — | (0.22, 0.44) | 100 | 98.9 |
| null | high | informative | 0.14 | (−0.06, 0.33) | 0.51 | (−1.03, 1.23) | 63.0 | 46.7 |
| large | low | vague | 0.58 | (0.35, 0.86) | 0.31 | (−0.19, 1.42) | 94.6 | 87.9 |
| large | low | none | 0.49 | (0.36, 0.62) | — | (0.36, 0.62) | 100 | 100 |
| large | low | informative | 0.59 | (0.34, 0.86) | 0.35 | (−0.19, 1.43) | 93.9 | 85.8 |
| large | high | vague | 0.71 | (0.55, 0.89) | 0.39 | (−0.17, 1.63) | 95.5 | 90.4 |
| large | high | none | 0.58 | (0.47, 0.69) | — | (0.47, 0.69) | 100 | 100 |
| large | high | informative | 0.72 | (0.55, 0.89) | 0.39 | (−0.10, 1.60) | 96.0 | 90.8 |

Note: *d*, summary effect estimate; *Prec(d)*, precision (1/variance) of summary effect estimate; *sd*, between-trial standard deviation; *CrI*, 95% credible interval; *d.new*, expected effect estimate in an individual (new) trial; *p1*, probability that $d.new \geq 0$; *p2*, probability that $d.new \geq 0.2$; for explanation see text and the Appendix.

small effect size ($d = 0.2$) can be expected with a probability of 86.1%. However, if prior beliefs are critical, say, LTPP is considered no better than any comparator psychotherapy (zero treatment effect) with high certainty (high precision) and realistic between-trial variance (informative standard deviation), the average treatment effect reduces to 0.14 with a 95% credible interval −0.06 to 0.33, and LTPP is expected to reach an at least small effect in a new trial only with a probability of 46.7%. However, convinced representatives of LTPP may expect a large effect with high precision and a vague prior for the between-trial variance [for example, because they acknowledge that LTPP is hard to standardize completely and represents an "umbrella concept" (Leichsenring and Rabung, 2011)], which leads to a Bayesian average effect size estimate of 0.71 and a 90.4% probability that LTPP will outperform any less intensive psychotherapy with an at least small effect size in a new trial or individual setting. It should also be noted that fixed-effect models that do not model between-trial heterogeneity provide the same narrow credible intervals for the average summary effect and the expected effect in a new trial, both excluding zero in all scenarios. This finding shows that ignoring heterogeneity is likely to lead to a high confidence with regard to the effect to expect in an individual setting. In summary, the presented conclusions differ to a large extent in dependence of the prior beliefs, although the same data were analysed. Several other combinations of prior beliefs can be found in Table 1, which is likely to give an impression how they influence Bayesian estimates.

## From lumping to splitting through investigation of heterogeneity

If heterogeneity is observed, reviewers have several options to deal with it (Higgins and Green, 2011). Besides some less favourable options like excluding outliers or simply ignoring heterogeneity by using a fixed-effect model (consequences in a Bayesian framework can be seen earlier), most experts agree that the sources of heterogeneity should be investigated (Thompson, 1994; Lau *et al.*, 1998; Thompson and Sharp, 1999; Higgins *et al.*, 2002; Higgins and Green, 2011). The first step in such an investigation is the identification of possible moderator variables, i.e. characteristics that may influence the treatment effect. Two frequently used methods for further proceeding are subgroup analysis and meta-regression (both can be adapted in a frequentist or a Bayesian framework). *Subgroup analysis* is a division of the total set of trials into groups with a subsequent meta-analysis in each group. Comparing summary effects between groups visually or statistically, e.g. through performing a standard test for heterogeneity *across* subgroup results (Borenstein *et al.*, 2008), provide some information whether the variable

chosen to build groups explains a piece of heterogeneity. Similarly to simple regression, in *meta-regression* the effect size estimate in each trial (outcome) is regressed on one or more possible effect moderators (predictors). Although several models are available, random-effects meta-regression with a restricted maximum-likelihood estimator may be preferable (Thompson and Sharp, 1999). As clinical heterogeneity is one of the possible sources of statistical heterogeneity, it is frequently advisable to consider clinical characteristics as possible effect moderators. It should be mentioned that although *post hoc* identification of sources of heterogeneity is sometimes helpful, potential effect moderators should be defined a priori to avoid data dredging whenever possible (Higgins and Green, 2011).

In their meta-analysis, Leichsenring and Rabung (2011) considered the number of included trials too small to test diagnostic classification of included patients and comparator treatments as possible effect moderators. However, being confronted with the wide clinical range of investigated patients and comparator treatments several readers may not perceive a reduction in clinical heterogeneity solely due to the small number of trials. That this question is considered relevant is reflected by several reactions on the first meta-analysis on LTPP by Leichsenring and Rabung (2008), which used similarly broad criteria (Beck, 2009; Kriston *et al.*, 2009; Bhar *et al.*, 2010). Accordingly, for at least some researchers it is of interest to test whether clinical heterogeneity may provide some explanation for the statistical heterogeneity detected. Although trustworthiness and power of tests to explain heterogeneity in a small number of studies is in fact limited, running two random-effects meta-regression analyses provides informative results. Clinical heterogeneity due to diversity regarding mental disorders of included patients explained 32% of the between-study variance of effect estimates ($P = 0.16$), while diversity due to the different comparators explained 70% of the between-study variance of effect estimates ($P = 0.003$) (see Appendix for details). Even if the results are statistically not strictly convincing, one should keep in mind that "the extent of statistical heterogeneity, which can be quantified, is more important than the evidence of its existence" (Thompson, 1994; p. 1352). In case of investigation of heterogeneity *description* should be given priority over *inference*, not only because an "absence of evidence is not evidence of absence" (Altman and Bland, 1995), but also because heterogeneity tests frequently lack power to detect "even a moderate degree of genuine heterogeneity" (Thompson, 1994). In summary, the diagnostic categories of the included patients and particularly the classes of comparator treatments seem to substantially contribute to statistical heterogeneity among trials on LTPP.

## Splitting as an a priori strategy

Imagine teams developing disorder-specific clinical practice guidelines and coming to the question of reviewing evidence on LTPP. They would probably not think of summarizing treatment effects over diagnostic categories but rather keep diagnostic groups separated from the beginning. Even if they would be ready to accept that less intensive psychotherapies may show only negligible difference in effectiveness and be considered homogenous, their analyses still would be divided in diagnostic groups a priori. For another example picture a health-policy decision-maker in psychosomatic medicine looking over several institutions with various implemented psychotherapeutic practices. He or she may be not interested in diagnostic subgroups as most patients are suffering from multiple disorders in his or her psychosomatic clinics, but still would require evidence on LTPP in comparison to well-defined treatments. As he or she would want to know which of the current practices are inferior to LTPP and thus candidates to be substituted by it, analyses stratified for comparator treatments a priori would be necessary.

Results for these two scenarios are presented in Table 2. For subgroups according to patient populations, summary effect estimates (Hedges's $d$) range very broadly from 0.01 for Cluster C personality disorders (one study with a statistically non-significant effect estimate) over 0.34 for eating disorders (two studies with a pooled effect not reaching statistical significance) and 0.40 for depressive disorders (two studies yielding a statistically significant effect) to 0.81 for borderline personality disorder (five studies with a statistically significant effect but substantial heterogeneity between studies). Subgroup analyses according to comparator treatments yielded even more broadly varying results. While no statistically significant effect in comparison to dialectic behavioural therapy was found ($d = 0.17$, one study), comparisons with treatment as usual resulted in a statistically significant ($P < 0.001$) pooled effect size of 1.15 (three studies). Estimates for effectiveness of LTPP compared to cognitive therapy, clinical management, and mixed treatments fell between these two extremes. Thus, pooled standardized mean differences in subgroups according to patient populations and comparator treatments have a range (defined as the difference between the smallest and the largest estimate) of 0.80 and 0.98, respectively; remember, Cohen interprets a standardized mean difference of 0.80 as a "large" effect (Cohen, 1988).


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

**Table 2.** Results of subgroup analyses according to disorder of included patients and comparator treatment

| | | Effect estimate | | Heterogeneity | | |
|---|---|---|---|---|---|---|
| | *n* | *d* (95% CI) | *p* | *Q* (df) | *p* | $I^2$ (%) |
| *Disorder subgroups* | | | | | | |
| Borderline personality disorder | 5 | 0.81 (0.34 to 1.27) | <0.001 | 15.70 (4) | 0.003 | 75 |
| Eating disorders | 2 | 0.34 (−0.07 to 0.75) | 0.10 | 0.72 (1) | 0.40 | 0 |
| Depressive disorders | 2 | 0.40 (0.21 to 0.58) | <0.001 | 0.67 (1) | 0.41 | 0 |
| Cluster C personality disorders | 1 | 0.01 (−0.54 to 0.56) | 0.97 | n.a. | n.a. | n.a. |
| *Comparator subgroups* | | | | | | |
| Cognitive (behavioural) therapy | 3 | 0.39 (0.05 to 0.73) | 0.03 | 2.67 (2) | 0.26 | 25 |
| Dialectical behavioural therapy | 1 | 0.17 (−0.26 to 0.60) | 0.44 | n.a. | n.a. | n.a. |
| Structured clinical management | 1 | 0.65 (0.30 to 1.00) | <0.001 | n.a. | n.a. | n.a. |
| Treatment as usual | 3 | 1.15 (0.57 to 1.73) | <0.001 | 4.56 (2) | 0.10 | 56 |
| Mixed treatments | 2 | 0.33 (0.13 to 0.53) | 0.001 | 0.25 (1) | 0.62 | 0 |

Note: *n*, number of studies; *d*, standardized mean difference (Hedges' *d*); CI, confidence interval; calculations were made with Review Manager 5 (The Nordic Cochrane Centre, The Cochrane Collaboration, Copenhagen); n.a., not available.

It is of importance that although Leichsenring and Rabung (2011) defined their goal as investigating whether LTPP is superior to *shorter or less intensive* psychotherapy, in some of the included trials the administered comparator treatment was *similar in duration* to the LTPP arm. Due to this conflict between primary aim and inclusion criteria an interesting situation emerges, in that some of the subgroup analyses displayed in Table 2 answer a different question than the one posed by the higher-order meta-analysis, of which they are part of. For example, both in the only trial performed in Cluster C personality disorders (Svartberg *et al.*, 2004) and the only trial comparing LTPP with dialectic behavioural therapy (Clarkin *et al.*, 2007) the comparator treatments were similar in duration to the administered LTPP. In neither case a statistically significant difference for the overall effectiveness outcome was found. In the sense of the meta-analytical question they are rather negative findings, as LTPP was not superior to shorter psychotherapeutic treatments. However, considering them on their own may lead to classifying them as positive trials, because they support LTPP in comparison to established treatments of similar duration by suggesting non-inferiority. Although in general it is not advisable to include studies in a meta-analysis that do not contribute to answering the primary research question, this example shows that extreme care is needed when subgroup analyses are performed. Substantial clinical heterogeneity may lead to an unseen modification of the research question depending on which perspective is taken.

Further splitting through cross-tabulating estimators according to patient populations and comparator treatments (i.e. considering *both* of them simultaneously) leads to a conclusion that a reviewer strictly adhering to the PICO (Participants, Interventions, Comparisons and Outcomes) schema (Higgins and Green, 2011) may draw from the re-analysed meta-analysis:

> LTPP was more effective than treatment as usual (in three studies) and structured clinical management (in one study) but was not more or less effective than dialectical behavioural therapy (in one study) in patients with borderline personality disorder. LTPP was not more or less effective than cognitive therapy (in one study) or mixed psychotherapeutic treatments (in one study) in patients with eating disorders. LTPP was more effective than cognitive-behavioural therapy (in one study) and mixed psychotherapeutic treatments (in one study) in depressive disorders. LTPP was not more or less effective than cognitive therapy (in one study) in patients with cluster C personality disorders. For all other mental disorders and comparators randomized evidence on effectiveness of LTPP is missing.

## Extreme splitting and doubts on meta-analysis

Following the argumentation of the presented study up to this point a practicing clinician may raise the question, whether we need the statistics at all. An interesting fact is

that it takes much more time to read and appraise the meta-analytic literature and associated published correspondence on LTPP (produced in the last few years starting with the first meta-analysis of Leichsenring and Rabung in 2008) than the 10 studies that were analysed. Also, the PICO-conform summary in the previous section comes almost completely down to the discussion of individual studies. As clinical heterogeneity will be inevitably present in almost all meta-analyses (Thompson, 1994), in some cases consulting the primary studies may be more useful than relying on a meta-analysis which summarized trials that one considers inhomogeneous. A systematic review does not always need to contain a meta-analysis (O'Rourke and Detsky, 1989; Higgins and Green, 2011). As randomized studies investigating effectiveness of psychotherapeutic treatments are likely to have at least as much in common with complex health services interventions (Craig *et al.*, 2008) as with pharmacological clinical trials, the so called "realist review" methodology may also present an alternative to mechanic meta-analysis (Pawson *et al.*, 2005).

## Conclusions

The present study aimed at highlighting the consequences of following different strategies with regard to clinical heterogeneity in meta-analysis using an illustrative example. Several approaches along with their assumptions and implications for interpretation were described starting from a global lumping strategy and ending up at doubts whether meta-analysis is meaningful at all in some cases.

The main conclusion of the present study is a known one: in case of summing up evidence "one answer is not always enough" (Lau *et al.*, 1998). However, while previous work mainly focused on *dealing* with heterogeneity, here also underlying *assumptions* were explored and some practical tools were demonstrated that may aid the *interpretation* of meta-analytical findings with clinical heterogeneity. All analyses that were performed here can be conducted without consulting the primary studies as long as the meta-analysis of interest is adequately reported. Thus, readers may reflect their own assumptions and choose the fitting approach. It is likely that no single approach will fit to any context, but rather certain contexts will require certain approaches.

The lumping versus splitting dichotomy defines a "dimension" of level of abstraction with a global perspective at the one end and specific clinical decision-making situations on the other. The lumping "literature synthesis" approach provides a concise average summary effect and may be appropriate to satisfy regulatory and public-health requirements (Lau *et al.*, 1998). Calculating a prediction interval in addition to the summary estimate and its precision describes the range of true treatment effects to expect in an individual setting and may inform decision-makers considering implementation into clinical practice (Riley *et al.*, 2011). Bayesian methods allow the explicit modelling of prior beliefs, to be extended to incorporate several other parameters and multiple treatments, and thus may provide setting-specific and clinically useful information (Lu and Ades, 2004; Ades *et al.*, 2005). Splitting the evidence-base into subgroups, e.g. according to the treated disorder, may be essential to disorder-specific clinical practice guideline developers driven by the question "what works for whom" (Egger *et al.*, 2001; Fonagy, 2010b). Those seeking for advice at the bedside may consult primary studies or rely on non-meta-analytic reviews. It should be noted that this "dimension" of level of abstraction can be further extended. Further lumping is possible in the form of overviews of reviews or so called "umbrella" reviews (Whitlock *et al.*, 2008; Ioannidis, 2009; Caldwell *et al.*, 2010; Higgins and Green, 2011). On the other hand, one may prefer to look at certain subgroups in clinical trials (Yusuf *et al.*, 1991).

In the case study on LTPP clinical heterogeneity due to variation in diagnoses of included patients and form of administered comparator treatments was focused. Further sources of heterogeneity, e.g. due to varying outcomes, study design, methodological quality, and LTPP/comparator intensity were not investigated here in detail but partly addressed in the meta-analysis of Leichsenring and Rabung (2011). It should be noted that the decision to focus on patient diagnoses and comparator duration is an arbitrary choice. Others may find exploration of heterogeneity due to treatment duration or other issues more important. Thus, the presented re-analyses should be considered as illustrations rather than a comprehensive assessment of the effectiveness of LTPP. For completeness it should be mentioned that an independent meta-analysis by Smit *et al.* (2012) addressing a similar (but not exactly the same) research question with fairly comparable methods came to other conclusions than Leichsenring and Rabung (2011). Exploring the reasons for this inconsistency would go beyond the scope of the present study, but it is worth highlighting that conflicting meta-analytical findings are not uncommon in medical and psychological research (Pladevall-Vila *et al.*, 1996; Linde and Willich, 2003; Vavken and Dorotka, 2009; Kriston *et al.*, 2011; Rouder and Morey, 2011; Goodyear-Smith *et al.*, 2012). In the present study it was shown that even the sole interpretation of the same data set can be largely diverging


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

according to the context of the interpreter. Therefore, if we consider how many other decisions have to be made during a meta-analytic study, conflicting results seem little surprising. It is probably time to realize that meta-analyses are complex studies including a series of decisions to be made and thus always leaving space for subjectivity and with it also critique. More than finding the one and only "right" way of meta-analysis a transparent and well-documented reporting practice is needed, as recommended for example by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher *et al.*, 2009).

The presented work has several limitations. At some points it is certainly short-sighted. However, further details on specific issues would have only recited existing and well-documented knowledge and would have not contributed to the main topic considerably. Also, some of the examples may seem awkward or unreal; they should be seen as schematic and simplified illustrations rather than clear-cut case studies. Some of the applied statistical methods are rather complex and arguable. In order to keep the findings transparent a detailed Appendix is attached that allows the recalculation of every presented result.

Interpretation of meta-analytic findings has been shown to depend on the perspective of the reader. This perspective, the choice of the level of abstraction, can be defined as a subjective decision on the nature and amount of clinical heterogeneity that is considered irrelevant. As Lipsey and Wilson (2001) put it, "the definition of what study findings are conceptually comparable for purposes of meta-analysis is often fixed only in the eye of the beholder. Findings that appear categorically different to one analyst may seem similar to another" (p. 3). For some aims, for example for that of the authors of the re-analysed LTPP meta-analysis, loss of clinical information through summarizing effects across different populations and comparators may not be judged essential. However, if substantial clinical information threatens to be lost, one may seek a splitting level and avoid lumping. For this several strategies were presented. In some situations one is more appropriate than the other, but none is likely to fulfil every aim and sometimes more than one may be needed.

Probably the strongest implication of the present study is that "producers" and "consumers" of meta-analyses are encouraged to reflect their own beliefs on the role of clinical heterogeneity in the context of the objective they pursue. The central question is "Do I consider clinical differences between the summarized trials with regard to the effectiveness of the investigated treatment as irrelevant?" If the answer is "yes", summary estimates can be trusted. However, if the focus lays on effectiveness in an individual setting, additional calculations (e.g. prediction interval) should be considered. If no clear answer can be given to the question posed, strategies to explain heterogeneity (e.g. subgroup analysis, meta-regression) can be applied. If the answer is "no", i.e. clinical heterogeneity is judged substantial, splitting (or other control of moderators) seems inevitable. Consequently, what can be judged is not the appropriateness of a meta-analysis *per se*, but rather the fit between the underlying assumptions, analysis strategies, and the interpretation made. In further consequence, several seemingly methodological discussions on meta-analyses are likely to be traced back to conflicting underlying assumptions as well.

Although statistical methods were used to empirically *support* the presented conclusions, the key message, that clinical heterogeneity offers several interpretations, cannot be *proven* by statistics alone. The choice between a lumping or splitting approach is frequently a question of the context, value system, and beliefs of the reviewers, as well as of the higher-order objective of a study, which we may summarize with the term "*paradigm*". Interpretation of any finding is hardly possible without a context in which the interpretation is applied. In the end we will have to accept that methods of meta-analysis cannot be completely unified and freedom of interpretation is always present depending on the paradigm followed.

## Disclosures and acknowledgements

This study would have not been possible without the clear report of the meta-analysis by Leichsenring and Rabung (2011), whose published data were used. Furthermore, the author is grateful to Sven Rabung, Lars P. Hölzel, Alessa von Wolff, and Martin Härter for several enriching discussions on the topic of the article as well as to four anonymous reviewers and Alessa von Wolff for providing helpful comments on the manuscript.

## Declaration of interest statement

The author is not clinically trained, but he reports that the majority of his superiors and colleagues are trained in behavioural therapy rather than in psychodynamic approaches. He worked together with one of the authors of the illustrative example study, Sven Rabung, in the University Medical Centre Hamburg-Eppendorf, Hamburg, Germany for three years.

The study was not externally funded. The author reports no financial relationships with commercial interests and he is not a member of any scientific or health-care policy organization with special interests in the superiority of a certain psychotherapeutic approach over others.

The author has no competing interests.

## References

Ades A.E., Lu G., Higgins J.P.T. (2005) The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*, **25**(6), 646–654, DOI: 10.1177/0272989X05282643

Altman D.G., Bland J.M. (1995) Absence of evidence is not evidence of absence. *British Medical Journal*, **311**(7003), 485, DOI: 10.1136/bmj.311.7003.485

Beck A.T. (2009) Analyzing effectiveness of long-term psychodynamic psychotherapy. *Journal of the American Medical Association*, **301**(9), 931. DOI: 10.1001/jama.2009.178

Bhar S.S., Thombs B.D., Pignotti M., Bassel M., Jewett L., Coyne J.C., Beck A.T. (2010) Is longer-term psychodynamic psychotherapy more effective than shorter-term therapies? Review and critique of the evidence. *Psychotherapy and Psychosomatics*, **79**(4), 208–216, DOI: 10.1159/000313689

Borenstein M., Hedges L.V., Higgins J.P.T. (2008) Introduction to Meta-analysis, Chichester, John Wiley & Sons.

Caldwell D.M., Welton N.J., Ades A.E. (2010) Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *Journal of Clinical Epidemiology*, **63**(8), 875–882, DOI: 10.1016/j.jclinepi.2009.08.025

Clarkin J.F., Levy K.N., Lenzenweger M.F., Kernberg O.F. (2007) Evaluating three treatments for borderline personality disorder: a multiwave study. *The American Journal of Psychiatry*, **164**(6), 922–928.

Cochran W.G. (1954) The combination of estimates from different experiments. *Biometrics*, **10**(1), 101–129, DOI: 10.2307/3001666

Cohen J. (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edition, Hillsdale, NJ, Lawrence Erlbaum Associates.

Counsell C. (1997) Formulating questions and locating primary studies for inclusion in systematic reviews. *Annals of Internal Medicine*, **127**(5), 380–387.

Craig P., Dieppe P., Macintyre S., Michie S., Nazareth I., Petticrew M. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*, **337**, a1655, DOI: 10.1136/bmj.a1655

DerSimonian R., Laird N. (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**(3), 177–188.

DerSimonian R., Kacker R. (2007) Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, **28**(2), 105–114, DOI: 10.1016/j.cct.2006.04.004

Egger M., Smith G.D., Altman D.G. (2001) Systematic Reviews in Health Care: Meta-analysis in Context, London, BMJ.

Fonagy P. (2010a) The changing shape of clinical practice: driven by science or by pragmatics? *Psychoanalytic Psychotherapy*, **24**(1), 22–43, DOI: 10.1080/02668731003590139

Fonagy P. (2010b) Psychotherapy research: do we know what works for whom? *The British Journal of Psychiatry*, **197**(2), 83–85, DOI: 10.1192/bjp.bp.110.079657

Goodyear-Smith F.A., van Driel M.L., Del Mar C., Arroll B. (2012) Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: a case study. *BMC Medical Research Methodology*, **12**(1), 76, DOI: 10.1186/1471-2288-12-76

Gøtzsche P.C. (2000) Why we need a broad perspective on meta-analysis. *British Medical Journal*, **321**(7261), 585–586, DOI: 10.1136/bmj.321.7261.585

Greenland S. (1994) Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, **140**(3), 290–296.

Higgins J.P.T., Green S. (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration. http://www.cochrane-handbook.org [March 2011]

Higgins J.P.T., Thompson S.G., Deeks J.J., Altman D.G. (2002) Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Services Research & Policy*, **7**(1), 51–61, DOI: 10.1258/1355819021927674

Higgins J.P.T., Thompson S.G., Deeks J.J., Altman D.G. (2003) Measuring inconsistency in meta-analyses. *British Medical Journal*, **327**(7414), 557–560, DOI: 10.1136/bmj.327.7414.557

Higgins J.P.T., Thompson S.G. Spiegelhalter D.J. (2009) A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A*, **172**(1), 137–159, DOI: 10.1111/j.1467-985X.2008.00552.x

Ioannidis J.P.A. (2009) Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *Canadian Medical Association Journal*, **181**(8): 488–493, DOI: 10.1503/cmaj.081086

Kriston L., Hölzel L., Härter M. (2009) Analyzing effectiveness of long-term psychodynamic psychotherapy. *Journal of the American Medical Association*, **301**(9), 930–931, DOI: 10.1001/jama.2009.178

Kriston L., von Wolff A., Hölzel L.P. (2011) Contradictory high-level evidence for new generation antidepressants. [German]. *Psychopharmakotherapie*, **18**(1), 35–37.

Lau J., Ioannidis J.P.A., Schmid C.H. (1998) Summing up evidence: one answer is not always enough. *Lancet*, **351**(9096), 123–127, DOI: 10.1016/S0140-6736(97)08468-7

Leichsenring F., Rabung S. (2008) Effectiveness of long-term psychodynamic psychotherapy. *Journal of the American Medical Association*, **300**(13), 1551–1565, DOI: 10.1001/jama.300.13.1551

Leichsenring F., Rabung S. (2011) Long-term psychodynamic psychotherapy in complex mental disorders: update of a meta-analysis. *The British Journal of Psychiatry*, **199**(1), 15–22, DOI: 10.1192/bjp.bp.110.082776

Linde K., Willich S.N. (2003) How objective are systematic reviews? Differences between reviews on complementary medicine. *Journal of the Royal Society of Medicine*, **96**(1), 17–22, DOI: 10.1258/jrsm.96.1.17

Lipsey M.W., Wilson D.B. (1993) The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist*, **48**(12), 1181–1209.

Lipsey M.W., Wilson D.B. (2001) Practical Meta-analysis, Thousand Oaks, CA, Sage Publications.

Littell J.H., Shlonsky A. (2010) Making sense of meta-analysis: a critique of "effectiveness of long-term psychodynamic psychotherapy". *Clinical Social Work Journal*, **39**(4), 340–346, DOI: 10.1007/s10615-010-0308-z

Lu G., Ades A.E. (2004) Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, **23**(20), 3105–3124, DOI: 10.1002/sim.1875

Lunn D.J., Thomas A., Best N., Spiegelhalter D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistical Computing*, **10**(4): 325–337, DOI: 10.1023/A:1008929526011

Moher D., Liberati A., Tetzlaff J., Altman D.G., the PRISMA Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, **151**(4), 264–269, DOI: 10.1059/0003-4819-151-4-200908180-00135

O'Rourke K., Detsky A.S. (1989) Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology*, **42**(10), 1021–1024, DOI: 10.1016/0895-4356(89)90168-6

Pawson R., Greenhalgh T., Harvey G., Walshe K. (2005) Realist review – a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, **10**(Suppl 1), 21–34, DOI: 10.1258/1355819054308530

Pladevall-Vila M., Delclos G.L., Varas C., Guyer H., Brugués-Tarradellas J., Anglada-Arisa A. (1996) Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *American Journal of Epidemiology*, **144**(1), 1–14.

Riley R.D., Higgins J.P.T., Deeks J.J. (2011) Interpretation of random effects meta-analyses. *British Medical Journal*, **342**, d549, DOI: 10.1136/bmj.d549

Rosenberg M.S., Adams D.C., Gurevitch J. (2000) MetaWin: Statistical Software for Meta-analysis, version 2.0, Sunderland, MA, Sinauer.

Rouder J.N., Morey R.D. (2011) A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, **18**(4), 682–689, DOI: 10.3758/s13423-011-0088-7

Rubin D.B. (1992) Meta-analysis: literature synthesis or effect-size surface estimation? *Journal of Educational and Behavioral Statistics*, **17**(4), 363–374, DOI: 10.3102/10769986017004363

Smit Y., Huibers M.J.H., Ioannidis J.P.A., van Dyck R., van Tilburg W., Arntz A. (2012) The effectiveness of long-term psychoanalytic psychotherapy – a meta-analysis of randomized controlled trials. *Clinical Psychology Review*, **32**(2), 81–92, DOI: 10.1016/j.cpr.2011.11.003

Smith M.L., Glass G.V., Miller T.I. (1980) The Benefits of Psychotherapy, Baltimore, MD, John Hopkins University Press.

Smith T.C., Spiegelhalter D.J., Thomas A. (1995) Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine*, **14**(24), 2685–2699, DOI: 10.1002/sim.4780142408

Smith G.D., Egger M., Phillips A.N. (1997) Meta-analysis. Beyond the grand mean? *British Medical Journal*, **315**(7122), 1610–1614, DOI: 10.1136/bmj.315.7122.1610

Sutton A.J., Abrams K.R. (2001) Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, **10**(4), 277–303, DOI: 10.1177/096228020101000404

Svartberg M., Stiles T.C., Seltzer M.H. (2004) Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for cluster C personality disorders. *The American Journal of Psychiatry*, **161**(5), 810–817, DOI: 10.1176/appi.ajp.161.5.810

Thombs B.D., Bassel M., Jewett L.R. (2009) Analyzing effectiveness of long-term psychodynamic psychotherapy. *Journal of the American Medical Association*, **301**(9), 930, DOI: 10.1001/jama.2009.177

Thompson S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, **309**(6965), 1351–1355, DOI: 10.1136/bmj.309.6965.1351

Thompson S.G., Sharp S.J. (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**(20), 2693–2708, DOI: 10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V

Vavken P., Dorotka R. (2009) A systematic review of conflicting meta-analyses in orthopaedic surgery. *Clinical Orthopaedics*, **467**(10), 2723–2735, DOI: 10.1007/s11999-009-0765-2

Wachter K.W., Straf M.L., National Research Council (US) Committee on National Statistics. (1990) The Future of Meta-analysis, New York, Russell Sage Foundation.

Whitlock E.P., Lin J.S., Chou R., Shekelle P., Robinson K.A. (2008) Using existing systematic reviews in complex systematic reviews. *Annals of Internal Medicine*, **148**(10), 776–782.

Yusuf S., Wittes J., Probstfield J., Tyroler H.A. (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association*, **266**(1), 93–98, DOI: 10.1001/jama.1991.03470010097038

## Appendix

### A. Data extraction from the publication by Leichsenring and Rabung (2011)

As only analysis of the outcome "overall effectiveness" was reported in sufficient detail for re-analysis, this outcome was used. Effect sizes of the included primary studies were extracted from the forest plot displayed in the article by Leichsenring and Rabung (2011, Figure 2). Standard errors of the effect sizes were estimated from 95% confidence intervals, reported also in the forest plot, using the formula $se = (ub - lb)/3.92$, where *se* refers to the standard error and *ub* and *lb* to the upper and lower bounds of the confidence interval, respectively.

The extracted data used for further analysis were

| study | es | se |
|---|---|---|
| Bachar 1999 | 0.58 | 0.35 |
| Bateman 1999 | 1.76 | 0.36 |
| Bateman 2009 | 0.65 | 0.18 |
| Clarkin 2007 | 0.17 | 0.22 |
| Dare 2001 | 0.21 | 0.26 |
| Gregory 2008 | 0.70 | 0.38 |
| Huber 2006 | 0.53 | 0.19 |
| Knecht 2008 | 0.35 | 0.11 |
| Korner 2006 | 1.00 | 0.28 |
| Svartberg 2004 | 0.01 | 0.28 |

where *es* refers to the effect size and *se* to the standard error of the effect size.

Note that these effect sizes are based on available data, i.e. intent-to-treat estimates were calculated when sufficiently reported in the primary study, otherwise per-protocol data were used (Leichsenring and Rabung, 2011).

### B. Statistical software

Several statistical software packages were used because of their individual strengths and in order to check for possible differences between them. In *Review Manager 5* (The Nordic Cochrane Centre, The Cochrane Collaboration, Copenhagen) the Generic Inverse Variance module was used to obtain summary effects and heterogeneity estimates. In *SPSS/PASW Statistics 18* (IBM Corp., Armonk, NY) summary and heterogeneity estimates were calculated and meta-regression was performed with the MeanEs and MetaReg macros written by David B. Wilson (available from http://mason.gmu.edu/~dwilsonb/ma.html). In *Stata 10* (StataCorp, College Station, TX) the metan command was used to perform meta-analysis. In *MetaWin 2* (Sinauer Associates, Sunderland, MA) summary estimates and heterogeneity statistics were obtained. *WinBUGS 1.4* (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml) was used to perform Bayesian meta-analysis via Monte Carlo Markov chain simulation (code presented later). WinBUGS was run from *R* (http://www.r-project.org/) with the package R2WinBUGS (originally written by Andrew Gelman; changes and packaged by Sibylle Sturtz and Uwe Ligges; available from http://cran.r-project.org/web/packages/R2WinBUGS/index.html). Some additional calculations were made in *Microsoft Excel 2000* (Microsoft Corporation, Redmond, WA).

### C. Summary estimates in random-effects meta-analysis and heterogeneity statistics

Random-effects meta-analyses were performed using the extracted data (DerSimonian and Laird, 1986). Estimates with default number of decimal places are

| software | es | 95%ci | Q | p | var |
|---|---|---|---|---|---|
| RevMan | 0.54 | (0.30, 0.78) | 24.74 | 0.003 | 0.09 |
| SPSS | 0.5421 | (0.3017, 0.7825) | 24.7350 | 0.0033 | 0.086511 |
| Stata | 0.542 | (0.302, 0.782) | 24.735 | 0.003 | 0.087 |
| MetaWin | 0.5422 | (0.2648, 0.8195) | 24.7398 | 0.00327 | 0.0865 |

where *es* refers to the summary effect size, *95%ci* to the 95% confidence interval of the summary effect size, *Q* to Cochran's *Q*, *p* to the significance of the heterogeneity test, and *var* to the between-trial variance estimate.

Note that estimates of the summary effect size, of Cochran's *Q*, its corresponding *p*-value, and the between-trial variance show agreement across software packages (small differences are probably due to rounding). However, MetaWin reports a marginally larger confidence interval than the other software. This is very similar to the interval Leichsenring and Rabung report in their publication (Table 1) as 0.26 to 0.83 (the minor difference is probably due to rounding). Why MetaWin estimates of the bounds of the confidence interval differ from other software was not investigated further in the present study. Leichsenring and Rabung report a confidence interval of 0.41 to 0.67 in their forest plot (Figure 2), which is considerably narrower than the interval reported in their Table 1 and any estimate provided by other software. This discrepancy could not be completely explained but this uncertainty is very unlikely to have any serious effect on the conclusions of the present study.

Note that Leichsenring and Rabung report a *Q* statistic of 11.72 in their Table 1. This is in large agreement with the recalculated estimate from a *random-effects* model in MetaWin (12.1844, the small difference is likely to have occurred due to rounding or due to different handling of missing values). However, this value is *not* Cochran's *Q*. The *Q* statistic can be calculated from the inverse-variance weights and effect sizes as

$$Q = \sum\left(w \times es^2\right) - \frac{\left[\sum(w \times es)\right]^2}{\sum w}$$

where *w* refers to the inverse-variance study weight and *es* to the effect size in each study. Cochran's *Q* is calculated from *fixed-effect* study weights: $w = 1/se^2$, where *se* refers to the standard error of the effect size. This estimate can be obtained in MetaWin using the fixed-effect option as stated in the MetaWin user's manual (Rosenberg *et al.*, 2000). This is the value reported in the earlier table showing high agreement with estimates from other software. Using the *random-effects* option in MetaWin determines the study weights as $w = 1/(se^2 + var)$, where *var* refers to the between-trial variance component. A *Q* computed


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

with these weights can be seen as a "generalized" $Q$ statistic. This is the statistic that Leichsenring and Rabung reported in their meta-analysis and erroneously used to calculate $I^2$. However, this statistic does not correspond with wide-spread heterogeneity testing standards and is therefore of limited use.

### D. Calculating the prediction interval for the pooled estimate

Assuming normally distributed random effects Riley *et al.* (2011) approximate the prediction interval of a summary estimate as

$$cibounds = es \pm t_{k-2} \times \sqrt{se^2 + \text{var}}$$

where *cibounds* refers to the bounds of the prediction interval, *es* to the summary treatment effect, *se* to the standard error of the summary effect, *var* to the between-trial variance, and $t_{k-2}$ to the $100(1-\alpha/2)$ percentile of the $t$ distribution with $k$ as the number of studies in the meta-analysis and $\alpha$ usually chosen as 0.05 to give a 95% prediction interval.

Using the values from Leichsenring and Rabung (2011; Table 1) with $es = 0.54$ and $se = 0.1454$ (calculated from the reported confidence interval), and imputing $var = 0.0865$ (MetaWin estimate, see earlier) yields a prediction interval of −0.22 to 1.30. This value is reported in the main text of the present study. Note that using the SPSS estimates $es = 0.5421$, $se = 0.1226$, and $var = 0.086511$ (see earlier) provides a prediction interval of −0.19 to 1.28, which is only marginally different.

### E. Bayesian meta-analysis

Prior distribution of the treatment effect $d$ was defined as a normal distribution with a mean $m$ and precision (1/variance) *prec* parameter as $d \sim N(m,prec)$. For a vague distribution $m$ and *prec* were given the values 0 and 0.0001, respectively. A large $d$ was defined as 0.8. *Prec* was given 100 for high and 10 for low precision. The between-trial precision was given a gamma distribution with the shape and scale parameters $\Gamma(0.001, 0.001)$ for vague priors and $\Gamma(1, 0.1)$ for informative priors. In fixed-effect models the between-trial standard deviation was not estimated (fixed at zero).

The used WinBUGS code:

```
model{
 for(i in 1:ns) {
  y[i] ~ dnorm(delta[i],prec[i])    #trial effect (normal distribution)
  #y[i] ~ dnorm(d,prec[i])          #for fixed-effect
  var[i] <- pow(se[i],2)           #variance of trial effect
  prec[i] <- 1/var[i]              #precision of trial effect
  delta[i] ~ dnorm(d,tau)          #distribution of trial-specific effects
  }
 d ~ dnorm(0,.0001)                #vague prior average tr. effect
 #d ~ dnorm(0.8,100)               #prior large precise effect
 #d ~ dnorm(0.8,10)                #prior large imprecise effect
 tau ~ dgamma(.001,.001)           #vague gamma prior between-tr. precision
 #tau ~ dgamma(1,.1)               #prior informative gamma
 sd <- 1/sqrt(tau)                 #between-trial SD
 dnew ~ dnorm(d, tau)              #predicted trial effect
 r1 <- step(dnew)                  #prob. tr. better
 r2 <- step(dnew-.2)               #prob. tr. at least .2 better
}
```

The parameters *d, sd, dnew, r1*, and *r2* were monitored for the estimates reported in Table 1 of the main text.

### F. Meta-regression analyses

Information on patient populations and comparison groups was extracted from the report by Leichsenring and Rabung (2011; Data Supplement) using binary indicator variables for disorders and comparators as displayed here:

| study | eat | dep | clusc | bor | CBT | cm | DBT | Mixed | TAU |
|---|---|---|---|---|---|---|---|---|---|
| Bachar 1999 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Bateman 1999 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Bateman 2009 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Clarkin 2007 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Dare 2001 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Gregory 2008 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Huber 2006 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Knecht 2008 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Korner 2006 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Svartberg 2004 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

**Legend**

study study ID (first author, year of publication)
eat disorder indicator: eating disorder
dep disorder indicator: depressive disorders
clusc disorder indicator: cluster C personality disorders
bor disorder indicator: borderline personality disorder (reference category in meta-regression)
CBT comparator indicator: cognitive (behavioural) therapy
cm comparator indicator: clinical management
DBT comparator indicator: dialectic behavioural therapy
Mixed comparator indicator: mixed treatments
TAU comparator indicator: treatment as usual (reference category in meta-regression)

A random-effects meta-regression with a restricted-information maximum-likelihood estimator was applied (Thompson and Sharp, 1999). The estimates provided by SPSS are listed here:

According to disorder

```
    ------- Descriptives -------
  Mean ES      R-Square         k
   ,5223        ,3166       10,0000
      ------- Homogeneity Analysis -------
                    Q           df           p
Model            5,2120      3,0000        ,1569
Residual        11,2492      6,0000        ,0810
Total           16,4612      9,0000        ,0579
                          ------- Regression Coefficients -------
                    B           SE        −95% CI     +95% CI        Z           P        Beta
Constant         ,7468        ,1493        ,4541      1,0395       5,0011      ,0000       ,0000
eat             -,3894        ,2951       -,9678       ,1890      −1,3195      ,1870      -,3471
dep             -,3236        ,2308       -,7761       ,1288      −1,4021      ,1609      -,3738
clusc           -,7368        ,3752      −1,4723      -,0014      −1,9636      ,0496      -,5050
          ------- Restricted Maximum Likelihood Random Effects Variance Component ---
v                    =          ,04010
se(v)                =          ,04104
```


Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013). DOI: 10.1002/mpr

According to comparator

------- Descriptives -------

| Mean ES | R-Square | k |
|---|---|---|
| ,4726 | ,6978 | 10,0000 |

------- Homogeneity Analysis -------

| | Q | df | p |
|---|---|---|---|
| Model | 17,2599 | 4,0000 | ,0017 |
| Residual | 7,4751 | 5,0000 | ,1876 |
| Total | 24,7350 | 9,0000 | ,0033 |

------- Regression Coefficients -------

| | B | SE | −95% CI | +95% CI | Z | P | Beta |
|---|---|---|---|---|---|---|---|
| Constant | 1,1382 | ,1911 | ,7638 | 1,5127 | 5,9576 | ,0000 | ,0000 |
| CBT | -,7362 | ,2389 | −1,2045 | -,2680 | −3,0819 | ,0021 | -,9139 |
| clm | -,4882 | ,2625 | −1,0027 | ,0263 | −1,8599 | ,0629 | -,5066 |
| DBT | -,9682 | ,2914 | −1,5393 | -,3971 | −3,3229 | ,0009 | -,8433 |
| Mixed | -,8095 | ,2163 | −1,2333 | -,3856 | −3,7432 | ,0002 | −1,2097 |

------- Restricted Maximum Likelihood Random Effects Variance Component -------

| | | |
|---|---|---|
| v | = | ,00000 |
| se(v) | = | ,01436 |

Note: Categorizing studies according to mental disorders and comparators may seem somewhat arbitrary, even though the terminology reported by Leichsenring and Rabung (2011) was used. Nevertheless, changing any of these decisions is unlikely to influence the conclusions considerably. Furthermore, performing several statistical tests (e.g. in subgroup analyses) increases the probability of false-positive findings and fitting complex statistical models (e.g. multiple random-effects meta-regression) in a limited number of studies is challenging. Also, to support the conclusions more strongly, disorder indicators, comparator indicators, and their interaction terms should have been entered in a single model. Unfortunately this would have led to extreme problems with degrees of freedom and multicollinearity and was therefore not feasible.