International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(2): 166–174 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1381

Critical evaluation of mixed treatment comparison meta-analyses using examples assessing antidepressants and opioid detoxification treatments

ALEXANDER SCHACHT,[1] YULIA DYACHKOVA[2] & RICHARD JAMES WALTON[3]

1 Eli Lilly, Global Statistical Sciences, Bad Homburg, Germany
2 Eli Lilly Regional Operations, Ges.m.b.H., Vienna, Austria
3 Save Sight Institute, University of Sydney, Sydney Eye Hospital, Sydney, NSW, Australia

Key words
mixed treatment comparison (MTC), multiple treatment meta-analysis, inconsistency, incoherence, bias

Correspondence
Alexander Schacht, Lilly Deutschland GmbH, Werner-Reimers-Str. 2–4, 61350 Bad Homburg, Germany.
Telephone (+49) 6172 273 2728
Fax (+49) 6172 273 2182
Email: schacht_alexander@lilly.com

Received 17 January 2012;
revised 27 April 2012;
accepted 19 June 2012

Abstract

Comparing multiple treatment options using meta-analytical methods requires complex statistical methods called mixed treatment comparisons (MTCs). Such methods offer the possibility to summarize data from many clinical trials comparing the different available options. However, those methods are based on a number of assumptions and inherent difficulties that are discussed and illustrated with examples from the psychiatric literature to help readers to understand the strengths and weaknesses of these methods. This review will help enable readers to critically appraise the methodology and results of publications that use MTCs. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

For most diseases, physicians have the opportunity to choose amongst a variety of different treatment options in order to select the most appropriate treatment, which necessitates relative assessment. Additionally, reimbursement institutions and advisory health technology assessment institutes such as the National Institute for Health and Clinical Excellence (NICE) in the UK, the Transparency Committee in France, or the Institute for Quality and Efficiency in Health Care (IQWiQ) in Germany are concerned with evaluating relative benefits of treatments. Furthermore, academic institutions like the Cochrane Collaboration are reviewing literature to compare different treatments. Ordinary meta-analysis comparing two treatments is widely used and has long been studied. The number of meta-analysis studies as well as the number of studies on meta-analysis methodology have increased considerably over the last two decades (Sutton and Higgins, 2008).

An indirect comparison of two treatments A and B is performed by using trials comparing A and C and trials comparing B and C, thus A and B are compared indirectly via treatment C. Methods that include indirect comparison are evolving and the number of publications using such indirect comparisons has increased (Song *et al.*, 2009). Statistical methods that use both direct and indirect comparisons are called mixed treatment comparisons (MTCs), network meta-analyses, or multiple treatment meta-analyses. Two recent examples of MTC applications are a comparison of 12 second-generation antidepressants (Cipriani *et al.*, 2009) and a comparison of four treatments used for opioid detoxification (Meader, 2010). These studies will be used to illustrate the following discussion.

MTCs offer a promising new way to compare treatments and thus to make decisions in situations where relevant information is scarce. However, analysts who choose to implement MTCs currently face a growing number of practical and methodological challenges. This critical evaluation will outline three layers where problems may arise. Firstly, any problems with the initial randomized clinical trials (RCTs) upon which the MTC is based are carried through to the subsequent meta-analysis (Higgins *et al.*, 2008). Although RCTs have limitations they are widely accepted as the gold standard for comparing treatments and thus those specific limitations will not be discussed here.

Secondly, Victor (1995) and Salanti *et al.* (2008) among others, have pointed out that in contrast to RCTs, meta-analyses are essentially observational and are therefore vulnerable to potential selection, dissemination, and publication biases in the sense that reported treatment differences may be systematically overestimated or underestimated. These issues are also applicable to MTCs; in fact they may even be harder to detect and guard against and thus will warrant discussion.

Thirdly, an MTC analysis requires assumptions beyond those necessary for regular meta-analyses. Some of these are connected to the *similarity* assumption for adjusted indirect comparison and the *consistency* or *coherence* assumption for the combination of direct and indirect evidence. It is essential for meta-analysts and physicians to fully understand and appreciate these underlying assumptions and their resulting limitations in order to sensibly apply and interpret MTCs appropriately; therefore these assumptions will be examined in detail. Unfortunately, insufficient knowledge and flawed applications are frequent. Song *et al.* (2008) evaluated publications using indirect comparisons and found methodological problems related to unclear understanding of underlying assumptions, inappropriate search and selection of relevant trials, use of inappropriate or flawed methods, lack of objective or validated methods to assess or improve trial similarity, and inadequate comparison or inappropriate combination of direct and indirect evidence.

Summaries of MTC examples

Meader (2010) presents a comparison of four treatments for opioid detoxification: methadone, buprenorphine, clonidine, and lofexidine. The publication is based on 23 RCTs; 20 of those reported data on completion of treatment and were included in the meta-analysis. For inclusion, these studies were required to be clearly described as randomized, include at least 10 patients 16 years of age or older, and have a treatment duration not exceeding 12 weeks. Publication dates of these studies spanned a period of 26 years from 1980 to 2006. The number of pairwise comparisons available from this set of studies ranged from one to eight with numbers presented in Figure 1, where numbers in parentheses indicate the numbers of excluded studies. Therefore, each treatment had been directly compared to all other treatments on at least one occasion.

The outcome measure chosen in the treatment comparison was the proportion of patients completing treatment. The paper presents conclusions based on both direct and indirect comparisons. The treatments were compared primarily on the basis of the odds ratio (OR). Additionally, the probability to have the highest relative treatment effect, i.e. a ranking of the treatments, was used. MTCs were based on the Bayesian approach using the Markov Chain Monte Carlo method for simulation. The authors used a random effects model, indicating that the treatment effect was expected to vary from one study to another.

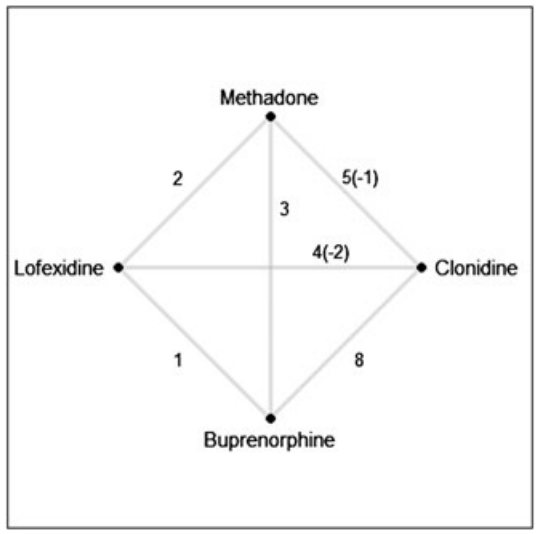

Figure 1. Comparison of four treatments.

All confidence intervals (CIs) estimated through utilization of the MTC technique were much narrower than those based on direct comparisons. This apparent increase in precision led to more statistically significant estimates of treatment effects.

Cipriani *et al.* (2009) present a comparison of 12 new-generation antidepressants (bupropion, citalopram, duloxetine, escitalopram, fluoxetine, fluvoxamine, milnacipran, mirtazapine, paroxetine, reboxetine, sertraline, and venlafaxine) as monotherapy in the acute-phase treatment of adults with major depression, excluding postpartum depression. The placebo arm was excluded from comparison, thus excluding from consideration two-arm studies with placebo versus active antidepressant as well as placebo treatment arms from studies with two or more antidepressants. The analysis was based on 117 studies of RCTs published between 1991 and 2007, with the number of direct comparisons ranging from six for milnacipran to 54 for fluoxetine. The network of comparisons is shown in Figure 2. More than a third of the direct comparisons were not available in the selected RCTs and a number of existing pairwise comparisons were based on only one study.

Two clinical outcomes were analyzed to compare treatments: "response" and "acceptability."

Response was defined as a patient having at least 50% improvement in the Hamilton Depression Rating Scale (HDRS) or the Montgomery–Åsberg Depression Rating Scale (MADRS) or a patient who scored "much improved" or "very much improved" on the Clinical Global Impression (CGI) scale at eight weeks. HDRS was given preference in cases where all three scales were reported and the timing of the response was extended from six to 12 weeks if eight-week results were not available. The term "acceptability" was used to describe the rate of study discontinuation for any reason during the first eight weeks (or 6–12 weeks if not available). All doses of medications were used in the primary analyses with sensitivity analyses based on more restricted doses that would be more comparable between different treatments.

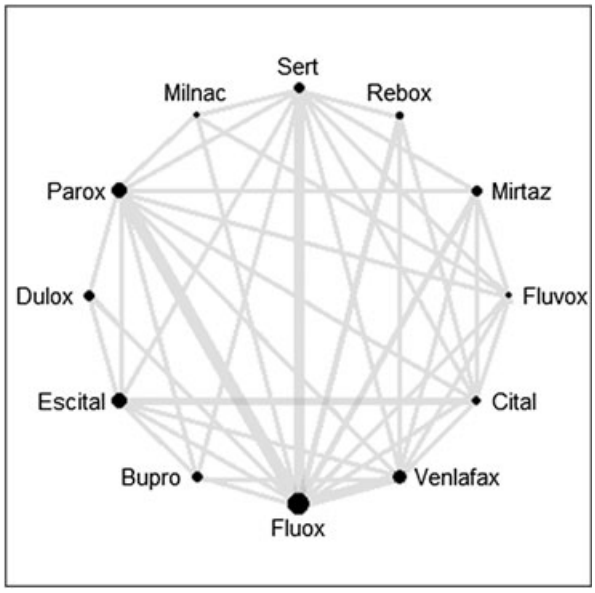

Figure 2. Network of comparisons.

"Moderate" heterogeneity and statistical incoherence in the study network were detected. Incoherence describes a situation whereby the estimates based on the direct comparisons are not contained in the corresponding 95% CI for indirect comparison estimates. Conclusions based on direct comparisons were much more conservative than those based on indirect ones.

Critical evaluation of meta-analyses in general

In order to appreciate the fundamental assumptions that underlie MTCs and to be able to critically appraise the results of MTCs, it is important to first identify those assumptions that already exist in the standard meta-analysis context which are also relevant in the MTC context (Victor, 1995). Furthermore, some of the challenges faced by any meta-analysis may be even more problematic for a network analysis using more studies and are thus discussed here.

Much discussion regarding meta-analysis revolves around terms like bias, homogeneity/heterogeneity, and other statistical terms that are often used interchangeably. Because there is often confusion around these terms, some basic statistical concepts need to be introduced first in order to assure correct understanding of such terminology. When deciding between two medications for a given patient, a physician is interested in the relative probabilities that the two medications will benefit the patient. The difference (sometimes computed as an OR) between these probabilities is the true treatment effect. As in any real-world scenario, the true treatment effect is unknown; however, under careful experimental conditions, it can be suitably estimated. RCTs provide estimates for this true treatment effect based on a sample of patients. As in all sample-based research, the estimates vary to a certain degree purely by chance arising from sample variation. However, other aspects of the experimental design may also systematically overestimate or underestimate the true treatment effect: this is known as *bias*. For example, as the parameter to be estimated is the average effect in a specified population, any deviation from this patient population might lead to bias. The full and fair assessment and reporting of potential biases in each contributing study is an important component of any meta-analysis. This task consumes significant amounts of time and valuable column inches in a published

Int. J. Methods Psychiatr. Res. 22(2): 166–174 (2013). DOI: 10.1002/mpr

manuscript. The current format and constraints of medical journals make it difficult to adequately report these mundane but crucial aspects of larger and more complex MTC analyses. Cipriani *et al.* (2009) mentions several different sources for bias, but a more in-depth review of all sources is surely beyond the capacity of a regular manuscript.

The presence of bias and heterogeneity is actually more the rule than the exception. Higgins (2008) has stated that *heterogeneity* is to be expected in a meta-analysis. He argues that it would be surprising if multiple studies, performed by different teams in different places with different methods, all resulted in estimating the same true treatment effect, while acknowledging that the actual effects measured of course will differ. It also needs to be mentioned that the definition of heterogeneity depends on the measure used to describe treatment effects, e.g. small heterogeneity for OR might translate into large heterogeneity for risk differences and vice versa. Thus, heterogeneity does not necessarily need to have a clinical background albeit this usually plays a role.

Important and interesting sources of heterogeneity are variables which have an influence on the treatment effect. These are sometimes called effect modifiers; they might be obvious and duly reported or they might not be easily identifiable. The latter case is especially challenging. Effect modifiers in psychiatry involve, but are not limited to, amount and flexibility of medication dosing, duration of treatment, pretreatment, inpatient versus outpatient setting, concomitant therapy, comorbidities, and other patient characteristics. Some commonly observed effect modifiers may also limit the validity of the MTC performed by Cipriani *et al.* (2009). When combining studies that were conducted over a long period of time, it becomes difficult to justify the assumption that all the factors remain comparable across an extended time period. This is because

- the standard of care may have changed,
- the diagnosis criteria may have further developed,
- treatment decisions (e.g. dosing) regarding a medication and practice have evolved,
- the quality of the implementation of the study (e.g. in terms of monitoring) may possibly have increased as more elaborate procedure and guidelines are developed,
- the countries participating in the studies may have changed.

The observed increase in placebo-response in studies evaluating antidepressants over time is probably a consequence of some of these changes (Rief *et al.*, 2009). Meader (2010) reports patients' mean age and the treatment length as possible effect modifiers and acknowledges that effect modifications might limit the validity of the MTC, but concludes that there was no evidence of such bias. Unfortunately, he does not indicate which results this conclusion is based upon. Also, given the brevity of the study characteristics reported, it remains difficult to assess the possible influence of other unreported variables being imbalanced over the studies.

To provide a reasonable overall estimate of the treatment effect, meta-analysis requires that the studies involved share the same study population, in the sense that baseline characteristics having an influence on the treatment effect (i.e. effect modifiers) are similar across the different study populations. It means that the study populations do not need to be the same with respect to all characteristics. Only those characteristics that might change the treatment response need to be consistent. This similarity of study populations is seldom or poorly discussed, as in Meader (2010). Cipriani *et al.* (2009) pooled the data from a phase IV study in patients attending UK primary care centers (McPartlin *et al.*, 1998) as well as a phase II dose-finding study including US outpatients in specialized investigational centers (Goldstein *et al.*, 2002).

Finally, if there are known treatment effect modifiers, i.e. treatment A might be better than treatment B for some patients and treatment B might be better than treatment A for other patients, it is obvious that it makes little sense to give every patient the same treatment. Inferences that one treatment is more efficacious than another that are derived from meta-analyses are conditional on the study population in question. Analyses including patient-level data on the treatment effect modifiers are needed and hence more detailed analyses on the RCT level or on individual patient data meta-analyses (or pooled analyses) are required in order to completely understand the relative benefit of treatment.

Pre-specification of the meta-analysis and posting of the protocol online before conducting the meta-analysis are increasingly accepted practices. This pre-specification is thought to protect against data-driven analyses and decision-making in a similar way as posting the protocol of a clinical trial on the Internet. It is well known and widely accepted that data-driven analyses can only be exploratory and not confirmatory. This is a higher hurdle for meta-analysis than for clinical trials, because the analyses usually incorporated in meta-analyses are more complex overall than those used in clinical studies. Thus, meta-analyses are not easy to conduct and require both statistical as well as clinical expertise in the relevant therapeutic field. Clinical expertise, in part defined by good knowledge of the literature, will however always lead to prior knowledge of relevant parts of the data used in the meta-analyses. Consequently, there cannot be any meaningful well-conducted meta-analysis in which the authors have no prior knowledge of the data. Of course, the

authors of a meta-analysis will not be familiar with every detail, but a general overview of the data can be expected. Hence, no decisions made concerning the analysis can be considered fully independent of the data. Therefore (to a certain extent) analyses decisions are data driven and thus the results of the analyses are exploratory. This is also acknowledged by Sutton *et al.* (2008), who state that carrying out a meta-analysis entails certain subjective components. These include the selection of studies, the decision concerning whether differences between study populations might impact the treatment effects, and the complexity of the model-fitting exercise. An iterative approach is often necessitated by the data at hand, e.g. if there is a subset of studies that are highly homogenous it makes sense to pool only these studies. However, this is by definition a *post hoc* decision.

Finally, there is vast literature about different sources of *bias* (optimism bias, quality-related bias, small study bias, sponsor bias, etc.) in meta-analysis, some elements of which have been discussed earlier.

The attempt to condense all the available data e.g. regarding outcome, outcome duration, treatment dose, and controls, into few or even only one measure might lead to oversimplification and thus to findings that are clinically difficult to interpret.

Critical evaluation of network meta-analyses

In the past, there have been occasions where the treatment arms from different studies were compared as if they would have been studied in the same trial. Such naive or unadjusted methods should categorically not be used, as commonly stated in the more recent literature about MTCs (Sutton *et al.*, 2008). It generates evidence equivalent to that of observational studies and should be avoided in the analysis of data from randomized trials. This method does not warrant further discussion here.

The first problem in meta-analysis using indirect comparison arises with the choice of network. It must be determined which *nodes*, i.e. which treatments, the network should contain and hence are the basis for the evidence. Such decisions should be based on the research question being asked. For example, Cipriani *et al.* (2009) decided to exclude placebo as a comparator whereas a similar analysis by Gartlehner *et al.* (2008) included placebo arms. However, both studies excluded trycyclic antidepressants, which is a questionable decision that merits further consideration because these medications are still used in many countries (such as Germany). The inclusion of placebo may have lead to different conclusions in the Cipriani study, as there are many more placebo-controlled studies than head-to-head comparisons. Further subjective decisions are required as to whether or not to aggregate, for example, medications of the same class (e.g. selective serotonin reuptake inhibitors, SSRIs), doses of the same medication, or flexible and fixed dosing. The decision regarding the network is ultimately driven by clinical knowledge about the relevant treatments (prohibiting again specification of the analysis without prior knowledge about the data) as well as to the practicalities of coping with the complexity of the network and the burden of a literature search comprising large networks.

Similarly to normal meta-analyses, an assumption of *homogeneity* is required for an MTC. Bucher *et al.* (1997) states that the only requirement is that the magnitude of the treatment effect is constant across differences in the populations' baseline characteristics. Unfortunately, this assumption is never met as Higgins and Thompson (2002) state in the introduction of their paper: "A systematic review of studies addressing a common question will inevitably bring together material with an element of diversity. Studies will differ in design and conduct as well as in participants, interventions, exposures, or outcomes studied." This is quite understandable and it would not be important to report these baseline characteristics if they have no prognostic value. This also links to the previous discussion about the shortcomings of meta-analyses.

The main concern regarding indirect comparisons is that they may be subject to greater bias than direct comparisons (Higgins and Thompson, 2002; Senn, 2000). Some statistical methods exist that may be used to summarize indirect effects, but they do not resolve many of the problems. Both Caldwell *et al.* (2005) and Sutton *et al.* (2008) try to explain the *similarity* assumption in the following way. One way to conceptualize the idea of similarity is to imagine that all trials included in the network had examined all treatments studied in the network, but that in each trial results for all but two or three treatments had been randomly lost. The key assumption for the fixed effect analysis is that the relative effect of one treatment compared with another is the same across the entire set of trials. This means that the true OR comparing treatment A with treatment B in trials of A versus B is exactly the same as the true OR for A versus B in the A versus C, B versus C, and indeed E versus F trials, even though A and B were not included in those studies. In a random effects model, in which it is assumed that the OR in each trial are different but from a single common distribution; the assumption is that this common distribution is the same across all sets of trials. If this assumption is not tenable, the analysis might be invalid. The Cipriani paper (Cipriani *et al.*, 2009) serves as an example where this is probably not the case. As mentioned earlier, the study populations

Int. J. Methods Psychiatr. Res. 22(2): 166–174 (2013). DOI: 10.1002/mpr

differ in many ways; for example, the studies were conducted across two decades. This argument about old and new studies is also used as an example of limited similarity by Sutton *et al.* (2008).

Further, Song *et al.* (2000) states that in addition to the homogeneity assumption, a similarity assumption is required for adjusted indirect comparison, known as *trial similarity*. This means that trials are similar in terms of moderators of relative treatment effect. That is, for the indirect comparison of treatment A compared with B based on the common placebo control, the average relative effect estimated by placebo-controlled trials of A should be the same as that of patients in placebo-controlled trials of B. Trial similarity could be considered from two perspectives: clinical similarity and methodological similarity. Clinical similarity refers to similarity in patients' characteristics, interventions, settings, length of follow-up, and outcomes measured. Methodological similarity refers to aspects of trials associated with the risk of bias. Caldwell *et al.* (2005) comments on this, stating that this assumption is unlikely to be statistically verifiable and it seems reasonable to rely on expert clinical and epidemiological judgment. Similarly, Ades and Sutton (2006) comment that every new type of information that is incorporated into a model requires assumptions which may not always be possible to check. The task of model critique also becomes considerably more difficult. One requirement is that investigators must check whether the combination of disparate types of evidence is sound. Another is to determine whether the different evidential sources are providing consistent evidence about the relative treatment effects. Often the concepts of "*validation*" and "*calibration*" are used in the decision-making literature; such approaches usually require that the information be split into training and validation data sets. The treatments are compared using the training data and this comparison is checked using the validation data. However, Ades and Sutton (2006) propose that all data should be used *ad hoc* and that the appropriateness of the incorporation of the data should be a matter of clinical or scientific judgment. Both, validation and calibration as well as clinical and scientific judgment based on model goodness-of-fit statistics have inherent flexibility and thus may lead to results, that can be interpreted in different ways.

This leads to the assumption about *coherence*, which is also called consistency, transitivity, or additivity (Sutton *et al.*, 2008; Lumley, 2002). Incoherence occurs, for example, when $A > B$ (meaning treatment A is better than treatment B) but $A < C < B$, that is, the direct comparison and the indirect comparison lead to alternate conclusions. Both Meader (2010) and Cipriani *et al.* (2009) discuss this problem. Incoherence might be due to common problems of heterogeneity also seen in normal meta-analysis, but it could also be that these are all valid results reflecting different patient populations (Lumley, 2002). This issue is only relevant to MTCs because it requires that at least three treatments be involved. Of course it is debatable what "$<$" means and how to assess something like a "A similar to B." Overall, coherence is a strong requirement that the data themselves cannot fully validate (Lu and Ades, 2004). More recently, Dias *et al.* (2010) summarize methods to check the consistency between direct and indirect effects.

The coherence problem increases with the number of treatments involved. Ades and Sutton (2006) state that combining direct and indirect evidence in MTCs relies heavily on additivity of effects across what may be a wide range, particularly when a large number of treatments are being compared. They would therefore expect much greater sensitivity of measures of consistency to scale assumptions than are seen in meta-analyses involving only two treatments (Deeks, 2002). In comparing the analyses by Meader (2010) with four treatments and the analyses by Cipriani *et al.* (2009) using 12 treatments, it is obvious that Cipriani *et al.* needs considerably more space to discuss this topic than Meader.

A further problem with incoherence is the low power often involved. Unfortunately, neither Meader nor Cipriani *et al.* discuss this issue. Salanti *et al.* (2008) state that although there is no literature addressing it, power will inevitably be low to detect inconsistency in a network. Thus, lack of demonstrable inconsistency in a MTC meta-analysis does not prove that the results of all trials are free of bias and diversity.

Salanti *et al.* (2008) allow inconsistency terms, but it is then unclear how comparisons between treatments can be made based on the combination of direct and indirect comparisons if they are not consistent. Salanti *et al.* (2008) argue that there is a trade-off to be made between the strength of consistency assumptions and the potential gain in precision achieved by combining direct with indirect evidence. They conclude that treatment-effect estimates from inconsistency models are difficult to interpret because inconsistency is a property of loops of evidence rather than of specific treatment contrasts. Baker and Kramer (2002) and Chou *et al.* (2006) present several examples in which the transitivity (coherence) hypothesis may not be a reasonable assumption to make due to differences in patient or trial characteristics; apparently, this is often seen as a result of some sort of bias. However, Sutton *et al.* (2008) comment differently. As the direct and the indirect treatment differences may arise from different

populations being studied, it could be wrong to call them biased. Rather, they may simply be inconsistent estimates. Song *et al.* (2000) conclude that the combination of inconsistent evidence from different sources may provide invalid or misleading results.

The combination of direct and indirect comparisons and thus of evidence with probably different *levels of validity* leads to another problem. The ratio of evidence from direct and indirect comparisons remains unclear; it seems that the weights depend on the number of patients included. In the Cipriani example, fluoxetine gets highly weighted for the direct effects versus the indirect effects because it is included in many studies, whereas duloxetine is weighted much more strongly on the indirect effect because it is a new compound with fewer clinical trials available to be analyzed. As the validity of indirect comparisons is less than that of direct comparisons, the validity of the ranking for duloxetine is thus reduced compared to the validity for fluoxetine. Though this is not explained, all treatments are ranked as if the same information with the same validity is available for all treatments. Recently, Dias *et al.* (2010) have proposed methods to distinguish between direct and indirect evidence in order to address this issue.

Some argue that MTC preserves *randomization*, but this claim seems too strong for a method that combines indirect and direct evidence across studies. Adjusted MTC takes the randomization into account but does not preserve it. The Cochrane Collaboration's guidance to authors states that indirect comparisons are not randomized, but are "observational studies across trials, and may suffer the biases of observational studies, for example confounding" (Higgins and Green, 2005). Salanti *et al.* (2008) further note that MTC meta-analyses should always be considered as retrospective observational investigations, even if they incorporate high-quality RCTs.

Conclusions

We have critically evaluated MTCs using examples from psychiatry and have found some important shortcomings. Therefore, the statement by Bucher *et al.* (1997) and others (Lu and Ades, 2004; Higgins and Green, 2005) that direct comparisons of treatment should be sought foremost is still valid. Direct randomized comparison is the most reliable way of comparing treatments (Lumley, 2002). When direct comparisons are unavailable, indirect comparison meta-analysis should evaluate the magnitude of treatment effects across studies, taking into account the limited strength of inference. Furthermore, when both indirect and direct comparisons are available, it is recommended that the two approaches are considered separately. The direct comparisons should take precedence in forming the basis for drawing conclusions (Higgins and Green, 2005).

Some difficulties arise if the direct evidence is inconclusive but the indirect evidence, either alone or in combination with the direct evidence, is not. Of course, head-to-head comparisons become increasingly impractical as the number of treatments increases. However, the validity of MTC-based conclusions is questionable and can lead to problems as well. Certainly, direct-comparison as well as head-to-head studies have methodological problems but they seem to be more transparent and easier to control. More transparency in MTCs and meta-analyses might be achieved by providing a separate "methods and motivation" publication. This will allow the documentation of the technical aspects of the analyses, e.g. the assessments of heterogeneity, inconsistency, and sensitivity analysis. The main publication simply does not have the space for this background material.

Another interesting point is made by Sutton *et al.* (2008). Sometimes, there may be one (or more) particular study in the meta-analysis that is more representative of the context for which the decision is being made. Therefore it should be expected that the treatment effect in practice is closer to that estimated from this specific study. The treatment effect estimated from the other studies in the meta-analysis might be irrelevant in this case.

Indirect comparisons may be a good source to plan the first direct-comparison study. But combining the indirect evidence with the outcome of that type of study in order to have more cumulative data again leads to the issues discussed earlier. The question is whether "all available evidence" or the "best available evidence" should be used (Song *et al.*, 2008). The assessment of data is worthwhile, but an opaque combination of all validity levels may not be helpful. Stepping down sequentially from the highest validity level to lower ones may be a better option. There is probably not one single "correct" way to combine data from related trials, but the "correct" analysis depends on the question of interest, which itself will depend directly on the decision question. This potentially conflicts with the notion of being able to produce one definitive summary of trial data, which would appear to be an aim of groups such as the Cochrane Collaboration (Sutton *et al.*, 2008).

The decision regarding the best way to combine data in a given situation needs the support of expert opinion. This is different than the position that expert opinion takes in the systematic review literature; in that case, in order to eliminate the selective use of data, the goal has been to make both the assembly of evidence and even its quality

Int. J. Methods Psychiatr. Res. 22(2): 166–174 (2013). DOI: 10.1002/mpr

assessment a repeatable, and thus mechanical, procedure (Song *et al.*, 2000). However, the burgeoning literature on different forms of bias in systematic reviews and meta-analyses (Song *et al.*, 2003; Juni *et al.*, 1995), that is perhaps a consequence of this mechanical approach, suggests that the strategy of limiting the influence of expert opinion has not been entirely successful (Song *et al.*, 2000).

In summary, indirect comparisons should be ranked, in terms of validity, below meta-analyses and head-to-head studies but may be on a similar level to observational data (or perhaps even below this, when the observational study is well-conducted) (Higgins and Thompson, 2002). Finally it should be noted that many of these comments apply to any MTC outside of psychiatry as well. Physicians should take these differences in the evidence level into account when deciding between treatment options.

In the absence of direct head-to-head comparisons, MTC methods can be helpful in decision-making when the decision-maker has a number of available treatment options and when the running of an RCT which includes all of the treatment options is impractical. As with any study, MTCs should follow good practice and guidance is available on how to conduct appropriate evidence gathering, meta-analysis methods, the completion of an appropriate network, and the synthesis of evidence. More research is required to understand the impact of some of the points raised in this paper. However, a well-conducted MTC study can provide the decision-maker with important information which would be otherwise unavailable in the existing literature.

Acknowledgements

All authors are affiliated with: Global Statistical Sciences, Eli Lilly and Company.

All authors have substantially contribution to the conception of the article, drafting the article and revising it critically for important intellectual content, and all gave final approval of the version to be published.

References

Ades A.E., Sutton A.J. (2006) Multiparameter evidence synthesis in epidemiologyand medical decision-making: current approaches. *Journal of the Royal Statistical Society Series A*, **169**(Part 1), 5–35.

Baker S.G., Kramer B.S. (2002) The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? *BMC Medical Research Methodology*, **2**(1), 13.

Bucher H.C., Guyatt G.H., Griffith L.E., Walter S. D. (1997) The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, **50**(6), 683–691.

Caldwell D.M., Ades A.E., Higgins J.P.T. (2005) Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *British Medical Journal*, **331**, 897–900.

Chou R., Fu R., Huffman L.H., Korthuis P.T. (2006) Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *Lancet*, **368**(9546), 1503–1515.

Cipriani A., Furukawa T.A., Salanti G., Geddes J.R., Higgins J.P., Churchill R., Watanabe N., Nakagawa A., Omori I.M., McGuire H., Tansella M., Barbui C. (2009) Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*, **373**(9691), 746–758.

Deeks J.J. (2002) Issues on the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, **21** (11), 1575–1600.

Dias S., Welton N.J., Caldwell D.M., Ades A.E. (2010) Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, **29**(7–8), 932–944.

Gartlehner G., Gaynes B.N., Hansen R.A., Thieda P., DeVeaugh-Geiss A., Krebs E.E., Moore C. G., Morgan L., Lohr K.N. (2008) Comparative benefits and harms of second-generation antidepressants: background paper for the American College of Physicians. *Annals of Internal Medicine*, **149**(10), 734–750.

Goldstein D.J., Mallinckrodt C., Lu Y., Demitrack M.A. (2002) Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial. *The Journal of Clinical Psychiatry*, **63** (3), 225–231.

Higgins J.P.T. (2008) Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, **37**(5), 1158–1160.

Higgins J.P.T, Green S. (eds) (2005) Cochrane handbook for systematic reviews of interventions. In Cochrane Library, Issue 2, Chichester, John Wiley & Sons.

Higgins J.P.T., Thompson S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**(11), 1539–1558.

Higgins J.P.T., White I.R., Wood A.M. (2008) Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials*, **5**(3), 2250–2239.

Juni P., Altman D., Egger M. (1995) Assessing the quality of randomized controlled trials. In Egger M., Smith G.D., Altman D.G. (eds) Systematic Reviews, pp. 87–108, London, British Medical Journal.

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

Lumley T. (2002) Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, **21**(16), 2313–2324.

McPartlin G.M., Reynolds A., Anderson C., Casoy J. (1998) A comparison of once-daily venlafaxine XR and paroxetine in depressed outpatients treated in general practice. *Primary Care Psychiatry*, **4**(3), 127–132.

Meader N. (2010) A comparison of methadone, buprenorphine and alpha2 adrenergic agonists for opioid detoxification: a mixed treatment comparison meta-analysis. *Drug and Alcohol Dependence*, **108**(1), 110–114.

Rief W., Nestoriuc Y., Weiss S., Welzel E., Barsky A., Hofmann S. (2009) Meta-analysis of the placebo response in antidepressant trials. *Journal of Affective Disorders*, **118**(1), 1–8.

Salanti G., Higgins J.P.T., Ades A.E., Ioannidis J.P.A. (2008) Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, **27**(2), 279–301.

Senn S.J. (2000) The many modes of meta. *Drug Information Journal*, **34**(2), 535–549.

Song F., Altman D.G., Glenny A.M., Deeks J.J. (2003) Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *British Medical Journal*, **326**(7387), 472–476.

Song F., Eastwood A.J., Gilbody S., Duley L., Sutton A.J. (2000) Publication and related biases. *Health Technology Assessment*, **4**(10), 1–115.

Song F., Harvey I., Lilford R. (2008) Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *Journal of Clinical Epidemiology*, **61**(5), 455–463.

Song F., Loke Y.K., Walsh T., Glenny A.M., Eastwood A.J., Altman D.G. (2009) Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *British Medical Journal*, **338**(7700), b1147.

Sutton A., Ades A.E., Cooper N., Abrams K. (2008) Use of indirect and mixed treatment comparisons for technology assessment. *PharmacoEconomics*, **26**(9), 753–767.

Sutton A.J., Higgins J.P.T. (2008) Recent developments in meta-analysis. *Statistics in Medicine*, **27**(5), 625–50.

Victor N. (1995) The challenge of meta-analysis: discussion. Indications and contra-indications for meta-analysis. *Journal of Clinical Epidemiology*, **1**(1), 5–8.

Int. J. Methods Psychiatr. Res. 22(2): 166–174 (2013). DOI: 10.1002/mpr