ELSEVIER

Journal of Clinical Epidemiology ■ (2009) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency

Deborah M. Caldwell*, Nicky J. Welton, A.E. Ades

Department of Community Based Medicine, University of Bristol, Cotham House, Cotham Hill, Bristol BS6 6JL, UK

Accepted 11 August 2009

Abstract

Objectives: To propose methods for mixed treatment comparisons (MTC) based on pooled summaries of the type produced in overviews of reviews.

Study Design and Setting: Overviews of reviews (umbrella reviews) summarize the results of multiple systematic reviews into a single document. They report the summary estimates from the original pairwise meta-analyses and discuss them in narrative form, with the intention of identifying the most effective treatment. We present methods for MTC synthesis, tailored for use with overviews of reviews. These generate a single internally consistent summary of all the relative treatment effects and assessments of whether the summary is consistent with the data. These methods are applied to a published overview of treatments for childhood nocturnal enuresis. We apply the methods to both fixed-effect (FE) and random-effects (RE) meta-analyses of the original trials.

Results: The summary relative risks based on FE meta-analyses, as originally published, were highly inconsistent. Those based on RE meta-analyses were consistent and could, given standard assumptions on comparability of treatment effects in meta-analysis, form a basis for coherent decision making.

Conclusion: Along with the summaries from systematic reviews, MTC methods should be used in overviews to provide a single coherent analysis of all treatment comparisons and to check for evidence consistency.

Keywords: Meta-analysis; Indirect comparisons; Network meta-analysis; Inconsistency; Nocturnal enuresis; Cochrane Collaboration

1. Introduction

Fifteen years ago the explosion in the number of randomized controlled trials created a need for the *systematic* review and synthesis of evidence of intervention effectiveness [1]. More recently, this has been matched by an explosion in the number of systematic reviews and a proliferation of treatment options. There are currently over 3,000 published reviews indexed on the Cochrane Database of Systematic Reviews, many of which can be considered multiple reviews of competing treatments for a single clinical condition. For example, consider the 22 reviews of interventions for adult smoking cessation, with approximately 42 distinct treatment regimes analyzed in 38 separate meta-analyses for the single outcome of abstinence at 6 months or more. This is not an isolated case, further notable examples are found in the management of asthma for adults (19 reviews) and management of primary hypertension in adults (9 systematic reviews indexed and a further 10 protocols registered).

One response to the increasing volume of systematic reviews is the overview of reviews or umbrella review [2–4]. Overviews summarize the results of multiple systematic reviews addressing the effects of two or more interventions for the same clinical condition into an ‘‘accessible’’ document [5]. The intended audiences for Cochrane overviews are health care decision makers, who are approaching the Cochrane Library for an answer to the question ‘‘which treatment should I use for this condition?’’ [6]. Overviews do not aim to repeat or update the literature searches, eligibility assessment, quality assessment, or evidence synthesis from the reviews that are summarized. The Cochrane handbook suggests that, in most cases, overviews should simply extract the results as reported in the component systematic reviews and reformat them in tables or figures.

However, this simple reformatting of summary estimates into a table can make it difficult for the decision-maker to form a coherent judgment regarding which treatment

* Corresponding author. Tel.: +0117 3310631; fax: +0117 9546672
E-mail address: D.M.Caldwell@bristol.ac.uk (D.M. Caldwell).

0895-4356/10/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2009.08.025

What's new section?

Overviews of reviews summarise results of multiple systematic reviews addressing multiple treatments for a single condition.

Current approaches for overviews can make it difficult to form a coherent judgement regarding which treatment should be used.

A Mixed Treatment Comparison is a method for simultaneously comparing multiple treatments in a single meta-analysis.

If the evidence is inconsistent, an MTC analysis can be used to detect it; if it is not, MTC can provide a single coherent analysis of the relative efficacy of all treatments.

Extreme inconsistency was identified by an MTC analysis of an overview of treatments for nocturnal enuresis.

should be used. For example, consider the results reported in Table 1, reproduced from the first published Cochrane overview for childhood nocturnal enuresis [6]. This summarizes the findings from seven separate systematic reviews of 10 treatments [7−13] based on the outcome ‘‘failure to achieve 14 days consecutive dry nights.’’ It is not immediately clear from Table 1 how the authors reached their conclusion:

> ‘‘It appears that enuresis alarms are the most efficacious method…’’[6]

If one considers ‘‘no treatment’’-controlled comparisons only, then dry-bed training+alarm (DBT+alarm) is the most efficacious treatment (relative risk [RR] 0.17) and *not* enuresis alarm (RR 0.38). However, note from Fig. 1 there are also four direct, ‘‘active vs. active’’ comparisons for enuresis alarm (vs. desmopressin, imipramine, DBT, and psychological therapy), which generate *further* ‘‘indirect’’ evidence that has the potential to alter the ranking of treatments [14−16] but is not considered in the overview. Note that alarm reduces risk of failure in three of the comparisons; however, it is less effective than DBT (albeit with considerable uncertainty).

Table 1
Umbrella review of enuresis treatments: pooled RR (95% CI) as reported in the original Umbrella review

Control (X)	Treatment (Y)	RR	95% CI
No treatment	Enuresis alarm	0.38	0.33, 0.45
Enuresis alarm	DBT	1.33	0.79, 2.24
Enuresis alarm	Desmopressin	0.71	0.50, 0.99
Enuresis alarm	Imipramine	0.73	0.61, 0.88
Enuresis alarm	Psychological therapy	0.68	0.52, 0.90
No treatment	Psychological therapy	0.69	0.55, 0.85
No treatment	DBT	0.82	0.66, 1.02
No treatment	DBT+alarm	0.17	0.11, 0.28
No treatment	Diclofenac	0.52	0.38, 0.70
No treatment	Imipramine	0.77	0.72, 0.83

Abbreviations: RR, relative risk; CI, confidence interval; DBT, dry-bed training.

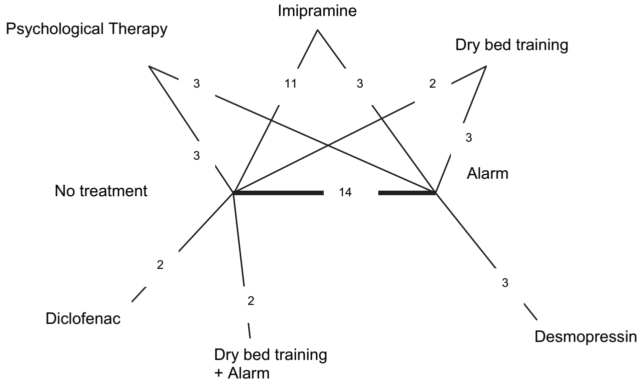

Fig. 1. Network of evidence for childhood nocturnal enuresis treatments. Each black line joining two treatments represents a direct head-to-head comparison.

Overviews can be considered mixed treatment evidence structures, that is, they can include both direct and indirect information on relative treatment effects. Statistical methods for analyzing multiple treatments simultaneously, in a single meta-analysis have been available for some time [17,18]. In this context, ‘‘indirect evidence’’ refers to evidence on treatment C relative to B obtained from A vs. B and A vs. C trials. A *mixed treatment comparison* (MTC) [19] or *network meta-analysis* [20] refers to ensembles of trial evidence in which direct and indirect evidence on relative treatment effects are pooled. The objective of an MTC analysis is to combine all the available trial evidence into an internally consistent set of estimates whilst respecting the randomization in the evidence. An MTC provides estimates of the effect of each intervention relative to every other, whether or not they have been directly compared in trials. One can also calculate the probability that each treatment is the most effective [21].

The new Cochrane handbook states that MTC analyses are ‘‘highly relevant’’ for overviews of reviews [22]. The purpose of this article is to propose an ‘‘aggregate’’ approach to MTC based on pooled summaries as reported in Cochrane overviews. We use the enuresis data in Table 1 to illustrate these methods and show how evidence consistency can be assessed in overviews of reviews. The article is structured as follows: we first outline the structure of the evidence network. We then describe an extension to the Bucher et al. [16] approach for checking evidence consistency. The ‘‘aggregate’’ MTC model is then described. Results are presented, followed by a discussion of possible sources of heterogeneity and inconsistency in overviews of reviews.

2. Method

2.1. Preparing the data

The enuresis network of evidence is shown in Fig. 1. The network is connected because there is a route (solid

Table 2
Pooled treatment effects and heterogeneity for all 10 pairwise comparisons reported in the umbrella review

			LRR (S.E.)		Heterogeneity (df)		
Control (X)	Treatment (Y)	RCTs	FE	RE	Cochran's Q	P-value	I^2 (%)
No treatment (1)	Enuresis alarm (2)	14	−0.97 (0.08)	−0.92 (0.16)	58.46 (13)	<0.0001	77.8
Enuresis alarm (2)	DBT (3)	3	−0.29 (0.26)	0.02 (2.06)	22.13 (2)	<0.0001	91.0
Enuresis alarm (2)	Desmopressin (4)	3	0.34 (0.17)	0.30 (0.17)	1.51 (2)	0.47	0.0
Enuresis alarm (2)	Imipramine (5)	3	0.31 (0.09)	0.53 (0.32)	6.02 (2)	0.05	66.8
Enuresis alarm (2)	Psychological therapy (6)	3	0.39 (0.14)	0.46 (0.46)	14.67 (2)	0.0007	86.4
No treatment (1)	Psychological therapy (6)	3	−0.37 (0.11)	−0.40 (0.43)	24.96 (2)	<0.0001	92.0
No treatment (1)	DBT (3)	2	−0.20 (0.11)	−0.19 (0.11)	0.04 (1)	0.85	0.0
No treatment (1)	DBT+alarm (7)	4	−1.77 (0.24)	−1.64 (0.70)	19.16 (3)	<0.0003	84.3
No treatment (1)	Diclofenac (8)	2	−0.65 (0.16)	−0.77 (0.56)	7.19 (1)	0.007	86.1
No treatment (1)	Imipramine (5)	11	−0.26 (0.04)	−0.38 (0.15)	476.63 (9)	<0.0001	98.1

Note: Estimates are shown on log RR scale and from both FE (reported in original overview—see text) and RE meta-analyses (derived from component systematic reviews—see text). Heterogeneity statistics (I^2 and Cochran's Q) and P-value as reported in the seven component systematic reviews. The standard errors are calculated from the confidence interval in Table 1.

Abbreviations: RCTs, randomized controlled trials; LRR (S.E.), log relative risk (standard error); FE, fixed effect; RE, random effects; df, degrees of freedom; DBT, dry-bed training; RR, relative risk.

lines) between each treatment and all the others. There are eight distinct treatments and 10 pairwise comparisons in the enuresis network. Treatments are numbered 1–8 and comparisons formed such that the comparator treatment has a lower numerical value than the experimental treatment. To ensure treatments were entered consistently it was necessary to use the inverse of the reported RR for four of the pairwise comparisons (enuresis alarm vs. DBT, enuresis alarm vs. desmopressin, alarm vs. imipramine, and alarm vs. psychological therapy). RR and 95% confidence intervals (95% CI) were log-transformed (log relative risk [LRR]) and standard errors calculated from the CIs reported in the overview (Table 2). The summary of RR cited in the original overview was based on fixed-effect (FE) meta-analyses. We obtained heterogeneity statistics (Q-statistic and I^2) from the seven component systematic reviews as they were not reported in the enuresis overview. Note (Table 2) that there is an evidence of substantial heterogeneity in 8 out of 10 of the comparisons. Consequently, we also conducted random-effects (RE) meta-analyses using the DerSimonian and Laird method in Stata for all 10 pairwise comparisons using the event data reported in the component systematic reviews.

2.2. *Checking for consistency across the enuresis network*

Suppose there are three treatments A, B and C, then d_{AB} is the relative effect of treatment B compared with treatment A, d_{BC} the relative effect of treatment C compared with treatment B, and d_{AC} is the relative effect of treatment C compared with treatment A. Then the fundamental assumption underpinning an indirect or MTC meta-analysis is that $d_{\mathrm{BC}}=d_{\mathrm{AC}}-d_{\mathrm{AB}}$. That is, the *true* underlying effect estimate of B vs. C is equal to the difference between A vs. C and A vs. B. Equations of this sort have been called "*consistency equations*" [23].

Bucher suggested a simple method for examining consistency in networks of three treatments. Suppose the data consist of three direct LRRs $\widehat{d}_{AB}^{Dir}$, $\widehat{d}_{AC}^{Dir}$, $\widehat{d}_{BC}^{Dir}$ and their variances $\mathrm{Var}_{\mathrm{AB}}^{\mathrm{Dir}}$, $\mathrm{Var}_{\mathrm{AC}}^{\mathrm{Dir}}$, $\mathrm{Var}_{\mathrm{BC}}^{\mathrm{Dir}}$. Note in Fig. 2 how these comparisons form a "loop" of evidence in which all three "edges" of the triangle are present. An indirect comparison of B vs. C can be formed $\widehat{d}_{BC}^{Ind}=\widehat{d}_{AC}^{Dir}-\widehat{d}_{AB}^{Dir}$ and the discrepancy between the direct and indirect estimates forms the measure of consistency $\widehat{\omega}_{\mathrm{BC}}=\widehat{d}_{BC}^{Dir}-\widehat{d}_{BC}^{Ind}$. The null hypothesis of evidence consistency ($\widehat{\omega}_{\mathrm{BC}}=0$) can then be tested by referring the test statistic z_{BC} to a standard normal distribution, where

$$z_{\mathrm{BC}}=\frac{\widehat{\omega}_{\mathrm{BC}}}{\sqrt{\mathrm{Var}(\widehat{\omega}_{\mathrm{BC}})}}$$

and

$$\mathrm{Var}\left(\widehat{\omega}_{\mathrm{BC}}\right)=\mathrm{Var}_{BC}^{Dir}+\mathrm{Var}_{BC}^{Ind}=\mathrm{Var}_{AB}^{Dir}+\mathrm{Var}_{AC}^{Dir}+\mathrm{Var}_{BC}^{Dir}$$

If the z_{BC} statistic leads us to reject the null hypothesis at say $P<0.05$, then we might conclude that there is

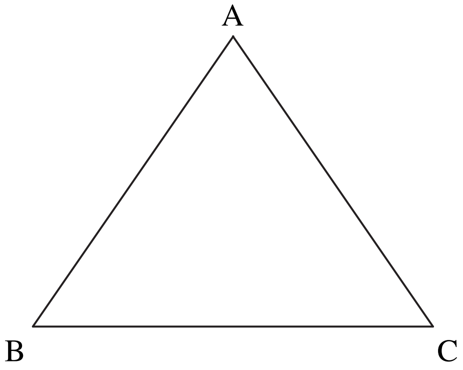

Fig. 2. A three-treatment network of evidence. All three possible "direct" comparisons have been made, as represented by solid black lines or "edges" of the triangle ("loop" of evidence).

inconsistency between the direct and indirect evidence. Note that in a three-treatment network only *one* measure of inconsistency is possible and inconsistency is a property of the "loop" of evidence A–B–C, rather than of the individual edges AB, AC, and BC [23].

2.3. Composite test for inconsistency

Note from Fig. 1 that there are three closed loops linking treatment comparisons where the "indirect" estimate can then be compared with the "direct" evidence, and thus three possible inconsistencies in the enuresis network;

- Alarm vs. no treatment—alarm vs. CBT—CBT vs. no treatment;
- Alarm vs. no treatment—imipramine vs. no treatment—alarm vs. imipramine;
- Alarm vs. no treatment—DBT vs. no treatment—alarm vs. DBT.

Gleser and Olkin [24], in a similar context, show how a χ^2 can be constructed to provide a composite test of the null hypothesis that evidence on all the contrasts is consistent. Their test is based on a weighted least squares regression and requires matrix algebra. The enuresis evidence network of Fig. 1 is, in fact, a special case in which there are four *independent* estimates of the same contrast (the three indirect comparisons of alarm vs. no treatment via psychological therapy, DBT, and imipramine and the direct comparison of alarm vs. no treatment). An approximate χ^2 test can therefore be constructed as follows. For M independent estimates, $\widehat{d}_m$, $m = 1,2\ldots$, of treatment effect each with a variance V_m, an estimate of the overall average is formed by inverse-variance weighting. With $w_m = 1/V_m$ a weighted average, $\tilde{d}$, of the M estimates is obtained:

$$\tilde{d} = \sum_{m=1}^{M} w_m \widehat{d}_m \Big/ \sum_{m=1}^{M} w_m$$

Then the following statistic can be referred to a χ^2 distribution, on M-1 degrees of freedom (df);

$$\sum_{m=1}^{M} w_m \left(\widehat{d}_m - \tilde{d}\right)^2 \sim \chi^2_{M-1}$$

2.4. Aggregate MTC method

The MTC model used here is based on the pooled summary estimates for each pairwise comparison and assumes consistency. The data inputs are the observed log relative risks LRR_{XY} and their variances σ^2_{XY}. The analysis was undertaken in a Bayesian framework and was implemented using WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/) [25]. The code is provided in Appendix 1. The approach is equivalent to Gleser and Olkin [24] but offers a simple solution in standard software, which is derived from earlier work on MTC [19,21].

The treatments are numbered 1, 2… k… NT, with Treatment 1 as the reference treatment. No treatment is used here as the reference treatment, although the choice is arbitrary provided the network is connected. The (NT-1) "basic" parameters represent the relative effect of treatment k to treatment 1 and were given vague prior distributions $d_k \sim N(0,100^2)$, $k = 2,3\ldots NT$. The effect of each treatment k relative to each treatment b is $d_{bk} = d_k - d_b$. A normal likelihood is assumed $LRR_{bk} \sim \text{Normal}(d_{bk}, \sigma^2_{bk})$. The goodness of fit of the model to the data can be measured by the posterior mean of the residual deviance [26]. The residual deviance is equal to the deviance for a given model, minus the deviance for a saturated model. For normally distributed data, the contribution to the deviance for each data point is given by,

$$Dev_{bk} = \frac{(LRR_{bk} - d_{bk})^2}{\sigma^2_{bk}}$$

where d_{bk} is based on the assumption of consistency. Because the likelihood is normal, the individual contributions to the residual deviance have a χ^2 distribution under the null hypothesis of consistency [27]. The sum of the individual deviance contributions provides an estimate of overall goodness of fit that can be referred to a χ^2 distribution with df equal to the number of data points. In a well-fitting model, the posterior mean of the summed deviance contributions $\overline{D}$ should be close to the number of data points, in this case 10.

Two "aggregate" MTC analyses were conducted, the first using the FE summary LRRs derived from the RR reported in the enuresis overview, and the second using RE summaries as input data. The programs were run for 50,000 iterations, 15,000 of which were discarded as burn-ins. Convergence was assessed using two chains and was achieved by 15,000 simulations based on the Brooks–Gelman–Rubin diagnostic tool in WinBUGS [28].

3. Results

3.1. Composite test of inconsistency

Using Bucher's measure of inconsistency, Table 3 compares the direct estimate of alarm vs. no treatment with each of the three indirect estimates, using (a) the summary LRRs from the pairwise FE meta-analyses and (b) summaries from pairwise RE analyses. Two of the three indirect estimates are significantly different to the direct estimate based on the FE summaries (Table 3a). Using the composite test, the χ^2 test of the null hypothesis of no difference between all four RR estimates gives a value of 18.8 on 3 df ($P = 0.0003$). This suggests that the enuresis evidence base is *highly* inconsistent using the FE summaries reported in the overview. Using the summary LRRs estimated from the RE analyses (Table 3b), the discrepancy is no longer statistically significant ($\chi^2 = 0.12$ on 3 df, $P = 0.98$).

Table 3
Comparison of the direct and indirect estimates (LRR) of the effect of enuresis alarms relative to no treatment

			Inconsistency (S.E.)		
Alarm vs. no treatment	$\widehat{d}$	S.E.	$\widehat{\omega}_{BC}=\widehat{d}_{BC}^{Direct}-\widehat{d}_{BC}^{Ind}$	z-statistic	P-value
a) Based on FE summary estimates from original overview					
Direct	−0.97	0.08			
Indirect via DBT	0.09	0.29	−1.06 (0.30)	−3.54	0.00
Indirect via psychological therapy	−0.58	0.10	−0.39 (0.13)	−3.06	0.00
Indirect via Imipramine	−0.76	0.18	−0.21 (0.20)	−1.08	0.28
Global test: $\chi^2 = 18.8$ (3df), $P = 0.0003$					
b) Based on RE summary estimates (as re-analyzed in Stata)					
Direct	−0.92	0.16			
Indirect via DBT	−0.23	2.04	−0.69 (2.04)	−0.34	0.74
Indirect via psychological therapy	−0.86	0.62	−0.06 (0.65)	−0.09	0.93
Indirect via Imipramine	−0.91	0.35	0 (0.38)	0	1
Global test: $\chi^2 = 0.12$ (3df), $P = 0.98$					

Abbreviations: LRR, log relative risk; S.E., standard error; FE, fixed effect; DBT, dry-bed training; RE, random effects.

3.2. MTC analysis of pooled summaries

Table 4 reports the 10 pairwise LRRs and standard deviations reported in the overview (observed) and compares them with the LRRs as estimated from the consistency MTC model (expected) for analyses based both on FE and RE summaries of the original trials. Note that for 3 of the 10 comparisons (asterisked in Table 4), that are not part of evidence loops in Fig. 1, the MTC estimate is identical to the observed summary.

In the FE summaries analysis, four treatment comparisons—alarm vs. no treatment, DBT vs. alarm, alarm vs. imipramine, and DBT vs. no treatment—show a discrepancy between the LRRs estimated in the MTC and those observed in the review. These are also picked out by the $\overline{D}$ statistic for ‘‘goodness of fit’’ of the model to the data (Table 4) [26]. In a well-fitting model, the total $\overline{D}$ contribution should be close to the number of data points, in this case 10. The summed $\overline{D}$ is 25.8, suggesting that the model using FE summaries is fitting poorly and that the assumption of consistency is not supported by the data.

Table 4 also reports the 10 LRRs and standard deviations estimated from pairwise RE meta-analyses and compares them with the LRRs estimated from the consistency MTC model (expected). The $\overline{D}$ is 7.1, suggesting that the model based on RE summaries could in principle form a coherent basis for decision-making, although there are further issues to be addressed (see discussion).

Table 5 gives the seven estimated RR and their credible intervals for each treatment relative to no treatment, and reports the probability (P[best]) each of the eight treatments is the most effective. Note P[best] are Bayesian P-values and their interpretation is analogous to a frequentist P-value. DBT+alarm has a probability of being the most effective treatment of 78%. The closest competitor is diclofenac with a 13% probability of being the most effective treatment for childhood nocturnal enuresis. Note that although the estimated relative efficacies should form the basis of decision-making, the probability [best] statistics, and the distribution of rankings on which they are based, provide considerable insight into the statistical strength of the differences between treatments, and perhaps represent an alternative to multiple testing. Similar statistics in a frequentist analysis could be generated by bootstrapping.

4. Discussion

Overviews of reviews have thus far been mainly narrative in form, presenting the summary estimates from the original pairwise meta-analyses in the form shown in the first three columns of Table 1, and without attempting a coherent, overall synthesis for a clinical decision maker. For multiple treatment evidence structures, a coherent answer to the treatment question *‘‘which treatment should I use for this condition?’’* requires a single statistical analysis providing internally consistent estimates for all possible pairwise treatment comparisons, based on both the direct and indirect evidence on each contrast, of the type generated by MTC synthesis [29]. Using simplified methods tailored to the summary relative effect data provided in overviews of reviews, we have shown that the evidence base, as originally reported in the overview, is highly inconsistent and does not form a coherent basis for decision-making. Moreover, when an internally consistent data ensemble is used, based on RE summaries, the original conclusion that DBT is most effective is thrown into doubt. Here DBT+alarm is most likely to be the effective treatment, but there is a 0.22 probability that it is not.

Our purpose here is to illustrate methods rather than to critique a particular overview, however, it is instructive to consider how to interpret our empirical findings of inconsistency in the FE summaries and apparent consistency with the RE summaries. Before we turn to this task, note that the assumptions made in a formal MTC analysis are also made, implicitly, in the original enuresis overview. To use

Table 4
Umbrella review of enuresis treatments: observed log relative risk (standard error) compared with posterior mean log relative risk from MTC analysis): fixed and random effects summaries.

			Fixed Effect Summaries			Random Effects Summaries		
Control (X)	Treatment (Y)	RCTs	Observed LRR LRR^{Dir}_{XY}, (SE)	Expected LRR d_{XY}, (SE)	Posterior mean Dev_{XY}	Observed LRR LRR^{Dir}_{XY}, (SE)	Expected LRR d_{XY}, (SE)	Posterior mean Dev_{XY}
No treatment (1)	Enuresis alarm (2)	14	-0.97 (0.08)	-0.78(0.06)	6.44	-0.92 (0.16)	-0.91 (0.14)	0.79
Enuresis alarm (2)	DBT (3)	3	-0.29 (0.26)	0.45 (0.11)	7.80	0.02 (2.06)	0.71 (0.18)	0.12
Enuresis alarm (2)	Desmopressin (4)	3	0.34 (0.17)	0.34 (0.17)	1.00	0.30 (0.17)	0.30 (0.17)	1.00
Enuresis alarm (2)	Imipramine (5)	3	0.31 (0.09)	0.49 (0.06)	3.85	0.53 (0.32)	0.53 (0.18)	0.31
Enuresis alarm (2)	Psychological therapy (6)	3	0.39 (0.14)	0.40 (0.09)	0.46	0.46 (0.46)	0.48 (0.32)	0.50
No treatment (1)	Psychological therapy (6)	3	-0.37 (0.11)	-0.38 (0.09)	0.66	-0.40 (0.43)	-0.43 (0.32)	0.55
No treatment (1)	DBT (3)	2	-0.20 (0.11)	-0.33 (0.10)	2.19	-0.19 (0.11)	-0.20 (0.11)	1.00
No treatment (1)	DBT + Alarm (7)	4	-1.77 (0.24)	-1.77 (0.24)	1.00	-1.64 (0.70)	-1.65 (0.70)	0.99
No treatment (1)	Diclofenac (8)	2	-0.65 (0.16)	-0.65 (0.16)	1.00	-0.77 (0.56)	-0.77 (0.56)	0.99
No treatment (1)	Imipramine (5)	11	-0.26 (0.04)	-0.29 (0.03)	1.43	-0.38 (0.15)	-0.38 (0.13)	0.84
$\overline{D}$					**25.83**			**7.09**

$\overline{D}$= posterior mean of summed deviance contributions. Dev = deviance

the data provided to answer the question ‘‘which treatment should I use for childhood nocturnal enuresis?’’ the authors must assume that the included trials represent a coherent body of data in which relative treatment effects are effectively identical or at least exchangeable throughout. The danger in the informal approach currently used in overviews of reviews is that a decision maker will form treatment recommendations based on an incoherent analysis of the available data, and more seriously will be unaware of extreme inconsistency in the evidence base.

4.1. Potential sources of inconsistency in overviews of reviews

A very clear result from our analysis is that the 10 summary contrasts in the original review, based on FE summaries, represent a highly inconsistent set of data, whereas the 10 summaries based on RE analysis appear consistent. The use of RE summaries would be entirely justified a priori, given the remarkable degree of heterogeneity reported in the original reviews (Table 2). The RE summaries appropriately reflect a greater degree of uncertainty in the estimates that serve as inputs to the MTC, with the consequence of a reduction in the extent of conflict between the different data sources. Unlike the FE analysis, the RE analysis could, therefore, potentially form a coherent basis for decision making although a number of concerns remain to be addressed.

Firstly, although the MTC analysis based on RE summaries provides an excellent fit to the data, the level of heterogeneity in the pairwise meta-analyses raises a series of questions about the robustness of the MTC findings. If the RE summaries are each a pooled average of highly heterogeneous trials it is legitimate to ask what exactly they are estimates of. The heterogeneity could be a consequence of ‘‘lumping’’ together disparate treatments or doses. Of particular concern here are the psychological therapy comparators that are described as ‘‘psychotherapy/cognitive/counseling’’ interventions and imipramine, for which dose information is not considered—it varies from 25 mg to 75 mg depending on the age of the child [13]. If the objective is to identify the most effective treatment, one needs to be able to specify each format of treatment, rather than estimate the average effect of what are in fact different treatments. Covariates, such as dose or type of psychological therapy can readily be incorporated into an MTC via meta-regression [30,31]. Another source of heterogeneity is, of course, the type of patient. For example, the severity of nocturnal enuresis and the response to treatment is likely to vary by age; a 16-year-old with nocturnal enuresis will require different treatment to a 5-year-old, or even a 10-year-old. We note that the participant eligibility criteria for the seven original systematic reviews relied on the trialists’ definition of ‘‘children.’’ In one review [8] trials of children aged 7−18 [32] and 8−14 [33] were combined with trials of those aged 4−14 [34]. We also note that some

Table 5
Posterior median RR (95% CrI) of each treatment relative to no treatment, and probability that each treatment is the most effective for outcome failure to achieve 14 days consecutive dry nights

Treatment	Probability (best)	RR relative to no treatment (95% CrI)
No treatment (1)	0	1
Alarm (2)	0.08	0.40 (0.31, 0.53)
DBT (3)	0	0.82 (0.66, 1.03)
Desmopressin (4)	0	0.54 (0.35, 0.84)
Imipramine (5)	0	0.68 (0.53, 0.89)
Psychological therapy (6)	0.01	0.65 (0.35, 1.22)
DBT+alarm (7)	0.78	0.19 (0.05, 0.76)
Diclofenac (8)	0.13	0.46 (0.16, 1.38)

Abbreviations: RR, relative risk; CrI, credible intervals; DBT, dry-bed training.

trials that included daytime enuresis and children with an organic cause of enuresis were included where the primary diagnosis was nocturnal enuresis. Both treatment “lumping” and patient heterogeneity weaken the validity and interpretability of the MTC estimates.

Whatever the origins of the evident heterogeneity, it points to the presence of unidentified treatment effect modifiers, and, particularly in view of the small numbers of trials for many of the treatment contrasts, this introduces the strong possibility of confounding between effect modifiers and treatment contrasts. A more fundamental concern is whether the body of trials in the overview represent a secure basis for a decision on “which treatment is best.” The treatments being compared are not likely to be suitable for all patients: some would be for younger children, whereas others would be more suitable for older, more severely affected patients who had failed on the simpler treatments. The high levels of heterogeneity among trials suggest that target populations for the treatment decisions need to be more tightly defined. Consequently, although DBT+alarm had the highest probability of being the most effective treatment in our analysis we do not suggest that this is a robust conclusion.

At a technical level, it may be worth noting that “overviews” can be vulnerable to biases that might generate *apparent* inconsistency. The meta-analytic methods used in the component systematic reviews should also be examined. Computational problems can arise in meta-analysis when observed risks are close to 0 or 1 and choice of continuity correction can bias the pooled summary [35]. Some frequently used meta-analytic methods can be seriously biased [36] and we note that 20 out of 35 trials in the enuresis network had zero cells. The scale of measurement used in the enuresis reviews also deserves reconsideration as the consistency equation underpinning MTC assumes additivity on the chosen scale [19]. The enuresis overview reported RR that are known to be problematic in meta-analysis because of the asymmetry introduced by “flipping” the coding of the event (p) to the non-event (1-p) [37]. The switching of events can have a substantial impact on reported effect size, its significance, and observed heterogeneity [38].

Finally, it is worth considering how the “aggregate” level MTC described here, based on contrast summaries from standard pairwise meta-analyses, compares to the “trial level” MTC analyses discussed in the literature to date. A thorough comparison is the subject of ongoing work, thus a detailed discussion of the advantages and disadvantages of both methods is beyond the scope of this article. Note that it is not our intention to propose aggregate MTC as a replacement for trial level MTC. Indeed, it is our view that trial level analysis is to be preferred for several reasons. Firstly, the overview represents a further level of aggregation, distancing the analyst still further from the original individual patient data, making it harder to control for confounders through meta-regression and other techniques. Aggregate MTC may therefore be more susceptible to the ecological fallacy [39]. Secondly, as noted above, the contrast summaries are often generated by various kinds of approximations. Thirdly, eight of the trials had multiple arms, which are reported in each of the constituent reviews without sample size adjustment. Although this is reasonable in the original pairwise reviews it causes problems in an overview as it involves a degree of “double-counting” that exaggerates the amount of data that is consistent, and hence leads to over estimation of consistency.

However, whilst trial-level MTC is to be preferred over aggregate methods, overviews of reviews reporting only pooled summaries of treatment effects are being increasingly produced both by the Cochrane Collaboration and elsewhere [3,40−42]. The methods proposed in this article provide a simple way of checking the consistency of the estimates assembled by these overviews and an approximate approach to identifying the most effective treatment.

Acknowledgments

This work was supported by grants from the Medical Research Council’s Health Services Research Collaboration. Part of this analysis was presented as Caldwell, D.M. and Ades, A.E. “Making sense of umbrella reviews.” *XV Cochrane Colloquium*, Sao Paulo, October 2007. We thank Prof. Paul Dieppe and Prof. Alex Sutton for helpful comments on an earlier draft of this article.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2009.08.025

References

[1] Mulrow CD. Rationale for systematic reviews. In: Chalmer I, Altman DG, editors. Systematic reviews. London: BMJ Publishing Group; 1995. p. 1−8.

[2] Becker LA, Oxman AD. Chapter 22: Overviews of reviews. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008.

[3] Cranney A, Guyatt G, Griffith L, Wells G, Tugwell P, Rosen C, et al. Summary of meta-analyses of therapies for post-menopausal osteoporosis. Endocr Rev 2002;23:570−8.

[4] Delgado-Rodriguez M. Systematic reviews of meta-analyses: applications and limitations. J Epidemiol Community Health 2006;60:90−2.

[5] Becker L. Umbrella reviews: what are they and do we need them?. Dublin, Ireland: XVI Cochrane Colloquium; 2006.

[6] Russell K, Kiddoo D. The Cochrane library and nocturnal enuresis: an umbrella review. Evidence-Based Child Health 2006;1:5−8.

[7] Glazener CMA, Evans JHC, Peto RE. Complex behavioural and educational interventions for nocturnal enuresis in children. Cochrane Database Syst Rev 2004;(1). Art. No.: CD004668. doi:10.1002/14651858.CD004668.

[8] Glazener CMA, Evans JHC, Peto RE. Alarm interventions for nocturnal enuresis in children. Cochrane Database Syst Rev 2005;(2). Art. No.: CD002911. doi:10.1002/14651858.CD002911.pub2.

[9] Glazener CMA, Evans JHC. Simple behavioural and physical interventions for nocturnal enuresis in children. Cochrane Database Syst Rev 2004;(2). Art. No.: CD003637. doi:10.1002/14651858.CD003637.pub2.

[10] Glazener CMA, Evans JHC, Peto RE. Drugs for nocturnal enuresis in children (other than desmopressin and tricyclics). Cochrane Database Syst Rev 2003;(4). Art. No.: CD002238. doi:10.1002/14651858.CD002238.

[11] Glazener CMA, Evans JHC, Cheuk DKL. Complementary and miscellaneous interventions for nocturnal enuresis in children. Cochrane Database Syst Rev 2005;(2). Art. No.: CD005230. doi:10.1002/14651858.CD005230.

[12] Glazener CMA, Evans JHC. Desmopressin for nocturnal enuresis in children. Cochrane Database Syst Rev 2002;(3). Art. No.: CD002112. doi:10.1002/14651858.CD002112.

[13] Glazener CMA, Evans JHC, Peto RE. Tricyclic and related drugs for nocturnal enuresis in children. Cochrane Database Syst Rev 2003;(3). Art. No.: CD002117. doi:10.1002/14651858.CD002117.

[14] Song F, Altman D, Glenny M-A, Deeks J. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. Brit Med J 2003;326:472−6.

[15] Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;26:1−148.

[16] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683−91.

[17] Eddy DM, Hasselblad V, Shachter R. Meta-analysis by the confidence profile method. London, UK: Academic Press; 1992.

[18] Hasselblad V. Meta-analysis of multi-treatment studies. Med Decis Making 1998;18:37−43.

[19] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105−24.

[20] Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313−24.

[21] Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Brit Med J 2005;331:897−900.

[22] Higgins JPT, Deeks J, Altman D. Chapter 16: Special topics in statistics. [updated September 2008]. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.0.1. The Cochrane Collaboration.

[23] Lu G, Ades AE. Assessing evidence consistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447−59.

[24] Gleser LJ, Olkin I. Stochastically dependent effect sizes. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York, NY: Russell Sage Foundation; 1994. p. 339−55.

[25] Spiegelhalter DJ, Thomas A, Best N, Lunn D. WinBUGS user manual: Version 1.4. Cambridge, UK: MRC Biostatistics Unit; 2001.

[26] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. JRSS (B) 2002;64:583−616.

[27] Dempster AP. The direct use of likelihood for significance testing. Stat Comput 1997;7:247−52.

[28] Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998;7:434−55.

[29] Salanti G, Ades AE, Higgins J, Ioannidis J. Evaluation of networks of randomised trials. Stat Methods Med Res 2008;17:279−301.

[30] Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. Am J Epidemiol 2009;169:1158−65.

[31] Nixon R, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. Stat Med 2007;26:1237−54.

[32] Sloop E, Kennedy W. Institutionalised retarded nocturnal enuretics treated by a conditioning technique. Am J Ment Defic 1973;77:717−21.

[33] Moffatt M, Kato C, Pless I. Improvements in self-concept after treatment of nocturnal enuresis: randomized controlled trial. J Pediatr 1987;110:647−52.

[34] Jehu D, Morgan R, Turner R, Jones A. A controlled trial of the treatment of nocturnal enuresis in residential homes for children. Behav Res Ther 1977;15:1−16.

[35] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351−75.

[36] Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analysis methods with rare events. Stat Med 2006;26:53−77.

[37] Cox DR. The analysis of binary data. London, UK: Methuen; 1970.

[38] Deeks JJ. Issues on the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575−600.

[39] Rothman KJ, Greeland S. Modern epidemiology. 2nd edition. Philadelphia, PA: Lippincott Williams & Wilkins; 1998.

[40] Stowe R, Ives1 N, Clarke CE, Deane K, van Hilten (sic), Wheatley K, et al. Evaluation of the efficacy and safety of adjunct treatment to levodopa therapy in Parkinson's disease patients with motor complications. Cochrane Database Syst Rev 2008;(2).

[41] Ryan R, Santesso N, Hill S, Kaufman C, Grimshaw J. Consumer-oriented interventions for evidence-based prescribing and medicine use: an overview of Cochrane reviews (protocol). Cochrane Database Syst Rev 2009;(2).

[42] Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, et al. Biologics for rheumatoid arthritis: an overview of Cochrane reviews (protocol). Cochrane Database Syst Rev 2009;(2).