Appendix L Additional details of methodology used

This appendix explains methods of clinical effectiveness analyses and syntheses in more detail and highlights any deviations from the Guidelines Manual (2012).

1 Data extraction

1.1 Time-points

The included evidence reported a variety of follow-up periods. In order to enable the comparison of studies with different follow-up periods, the GDG felt it important to extract outcomes at common time-points. Given the number and heterogeneity of the time-points reported in the literature, it was important to prioritise which time-points were extracted, while maintaining the ability to compare studies. Before data extraction commenced, the time-points where outcomes data were reported across the available literature were mapped and common time-points across the studies were chosen. The resulting time-points where outcomes were then extracted from the literature (when available) were as follows:

- 4 weeks +/- 7 days
- 8 weeks +/- 7 days
- 12 weeks +/- 14 days
- study end-point (if not one of the above).

If a study had more than one data-point available in this time period, the later time period was chosen. Rationale for including within 14 days for the 12-week time point, rather than within 7 days was to be able to include more data from more studies and possibly, more interventions (which was of particular interest due to the differential reporting of different studies). This meant that in some instances it was possible to include an outcome at 10 weeks and at 14 weeks in the same synthesis. Exploratory sensitivity

analysis suggested that this did not have a significant impact on the overall results from the analyses.

Extracting data for some time-points required the extraction of data from graphs, where numerical data were not available. This was typically performed only where dispersion was also available from the graph (for example, where error-bars were provided), and was performed by digitising the images in question and 'measuring' them with an electronic ruler.

A small number of studies reported some outcomes as averages of repeated measures over a period of follow-up, such as a mean change from baseline value to an average of measurements over weeks 2–12. In such cases, the data were recorded as reporting at the end of that period of follow-up (that is, at week 12 in this example).

1.2 Measurement tools extracted

A large number of different measurement tools were used in the literature for a number of critical and important outcomes (particularly global improvement, physical and emotional functioning and pain). The tools for which data were extracted were those prioritised by the GDG, based on clinical relevance, the reliability and validity of the tools for measuring particular outcomes, and the frequency to which they appeared in the literature.

Tools extracted	Tools not extracted
Patient-reported global pain: Patient-reported global impression of change (PGIC, 7-point)[1] Patient global impression of improvement (PGI-I)	Patient-reported global pain: Global assessment of therapeutic effect (GATE) Global pain relief (GPR) Global symptom score (GSS)
Physical and emotional functioning (including sleep)[2]: Brief pain inventory (BPI) - interference with function Beck's depression index (BDI) Centre for epidemiological studies-depression scale (CES-D) Hamilton rating scale for depression (HAMD) - 17 point Hamilton Depression Scale (HDS) (21-	Physical and emotional functioning (including sleep): Athens insomnia scale (AIS) Beck's anxiety index (BAI) Brief stress scale (BSS) Craig handicap assessment and reporting techniques (CHART) Expanded Disability status scale (EDSS) Functional independence measure (FIM) General Health Questionnaire (GHQ-12)

item version) Medical outcomes study sleep questionnaire (MOS) Profile of mood states (POMS) Hospital Anxiety & Depression Scale (Zigmund & Snaith) Euroqol - 5 dimensions (EQ-5D) Short form - 36 questions (SF-36)	Linear analog self-assessment scale (LASA) (NCCTG QoL) Minnesota Multiphasic Personality Inventory (MMPI) Zung pain and distress index (PAD) Pain catastrophising scale (PCS) Pain disability index (PDI) Sleep affective score (SAS) Self-assessment of treatment (SAT) Self-rating depression scale (Zung) (SDS) Spielberger State/Trait Anxiety Inventory (SSTAI/STAI) Sickness impact profile (SIP) Satisfaction with life scale (SWLS) West Haven-Yale multidimensional pain inventory (WHYMPI) pain intensity, life control, affective distress, interference with pain, social support, activity)
Pain: Numerical pain rating scale (NRS / NPRS) Neuropathic pain scale (NPS) Visual analogue scale for pain relief (VAS/VAS-PR) Visual analogue scale for pain intensity (VAS/VAS-PI) Visual rating scale (VRS) Brief pain inventory (BPI) McGill Pain Questionnaire (MPQ) / Short form McGill pain Questionnaire (SFMPQ) Sternback Pain intensity (SPI)	Pain: Oswestry Back Pain disability index (ODI) Pain-related self-statement scale (PRSS) Quality of life index (QLI)
	Others: Clinical global impression of change (CGIC) Clinical global impression of severity (CGI-S)
[1] PGIC was the primary tool for measuring patient-reported global change. This is the tool recommended by IMMPACT group (Dworkin et al. 2005) and is the tool most frequently used in the literature. [2] while outcomes related to physical and emotional functioning were extracted, the GDG felt it was inappropriate to synthesise results from different tools due to the variation in what different tools measured.	

1

2 Adverse effects: prioritising important events and approach to synthesis

The GDG was sent a questionnaire seeking the members' views on the 5 most important adverse effects to be considered for each drug class. The 6 top-rated outcomes for each drug type were extracted from each of the studies. Some GDG members listed additional adverse effects in the free-text section of the questionnaire. Since it was not possible to determine if other GDG members would prioritise these events if they had seen them listed in the questionnaire, these were also extracted.

To aid decision making, GDG advice was sought on those adverse effects judged to be clinically similar and therefore appropriate to be combined for analysis. In some cases it was not possible to combine adverse effects because they had been reported separately in a study, so to combine them would risk double counting. This is explained in table 1.

Table 1 Groupings of adverse effects for syntheses

Category	Comments
Blurred vision	None.
Burning pain	Includes studies that report 'burning sensation'
Cognitive impairment	Includes impaired attention and dissociation. The GDG felt these were similar enough to include under one larger subheading of ‘cognitive impairment’ (2 studies reporting 'mental change' were excluded as it was unclear what this meant).
Confusion	Not able to combine with cognitive impairment (because of duplicate reporting in some studies) so presented on its own.
Constipation	None.
Dizziness or vertigo	The GDG felt that combining dizziness and vertigo may be more useful to clinicians, because a diagnosis of vertigo requires determining certain physical features (which they thought were not likely to have been consistently applied in the studies) and that vertigo may be incorrectly categorised as dizziness. Six studies that reported ‘vertigo’ were excluded because they also reported dizziness separately. 'Balance disorder' was included in this synthesis but 3 studies reporting this were excluded because they reported both vertigo and dizziness separately.
Dry mouth	None
Fatigue or tiredness	Tiredness was considered similar enough to fatigue to combine both in one synthesis. Unable to be combined with lethargy (see below).
Gait disturbance	None.

Lethargy	While some GDG members felt this was similar enough to combine with fatigue and tiredness, it was not possible because some studies reported lethargy separately from fatigue (2 of the 4 studies that reported lethargy). It was necessary to exclude 1 study that reported 'lethargy and fatigue' as a combined outcome.
Mood disturbance	The GDG felt it was appropriate to include both depression and euphoria under a larger subheading of 'mood disturbance'.
Nausea	Nausea was reported in 58 studies and vomiting was reported in19; separate syntheses were done for nausea and for vomiting (3 studies reported 'nausea and vomiting' as one outcome but we felt it was inappropriate to include the results of these studies into either the nausea or vomiting syntheses).
Oedema	None.
Peripheral oedema	None.
Pruritus	None.
Rash / urticaria / overall erythema (not restricted to site)	The GDG felt it was acceptable to combine these general symptoms (differentiated from site-related).
Somnolence (including drowsiness and sedation)	Drowsiness, sedation and somnolence were considered similar enough to combine in one synthesis. One study reported 'daytime somnolence' and 'night-time somnolence': the rate reported for 'daytime somnolence' was included in this synthesis.
Urine retention	None.
Vomiting	As above under 'nausea'.
Weight gain	None.

There were 6 outcomes which were prioritised by the GDG as important but no evidence on these was identified in studies that met the inclusion/exclusion criteria:

- Addiction/dependence
- Anaphylaxis
- Disorientation
- Dyspnoea
- Hallucinations/paranoia/delusions
- Heaviness, pressure or tightness of any part of the body

The following were identified in free text by GDG members but no evidence was found in studies that met the inclusion/exclusion criteria:

- General neuroendocrine adverse effects

- Hormonal dysfunction
- Hyperalgesia
- Hyponatraemia
- Immunosupression
- Pituitary axis suppression
- Risk of transferring cream to eye or other sensitive parts of the body
- Sudden death (likely from cardiac arrhythmias)
- Tolerance
- Transient pain flare

Three adverse effects were extracted from the literature, but were not suitable for analysis. They are listed below with the reason why analysis was not undertaken:

- Postural hypotension was seen as important for antidepressants (reported in 2 studies: 1 comparing nortriptyline with gabapentin and another comparing amitriptyline with pregabalin; this network is not connected).
- Pro-arrhythmic effects for antidepressants (such as, arrhythmias, dysrhythmias, palpitation, and tachycardia) (reported in 5 studies: 2 showing there was no significant difference between amitriptyline and placebo, venlafaxine and placebo, and venlafaxine, imipramine and placebo, 2 showing insignificant differences between tramadol and placebo and amitriptyline and placebo) .
- Sexual dysfunction (2 studies showed no significant difference between treatments).

3 Crossover studies

Twenty-seven of the 116 included trials were crossover studies. The incorporation of data from RCTs of parallel and crossover design in single quantitative syntheses is a subject of methodological debate (see Elbourne et al. 2002). The following approaches were considered:

- The optimal method is to include data from crossover studies in a way that exploits the increased precision the crossover design provides. This is

straightforward where within-patient differences from a paired analysis are reported by authors; alternatively, methods are available that can impute these data if the correlation between treatment periods is known (or can be calculated) (see Elbourne et al. 2002). Unfortunately, very few of the included studies provided sufficient data to enable this approach to be adopted.

- Another method sometimes used is to restrict attention to the first period of randomised treatment in each crossover trial only. In this way, a parallel trial of half the size is derived. This approach is suboptimal, as it discards data from the remainder of the trial, and relies on data being reported in a way that facilitates the extraction of data from the initial period only. In the assembled evidence for this guideline, only 3 of 27 crossover studies reported the first treatment period separately. This approach was therefore rejected.
- An extreme option is to exclude all crossover studies from consideration. The GDG felt uncomfortable about this idea, particularly because the 27 studies constituted a large proportion of the head-to-head data available, and tended to be across a number of drugs and populations that were not covered elsewhere in the evidence base.
- Finally, it is possible simply to ignore the crossover design of the trials, and analyse them as if they had a parallel design. This method is not generally recommended, as it is ignores within-patient correlations and therefore discards the design advantages of crossover trials. However, this means that the approach is conservative, as it results in the trials having less weight in syntheses than they would have if paired data were used (or imputed). Therefore, when compared with the only practicable alternative of excluding crossover studies entirely (that is, giving them no weight at all), this was clearly a superior option.

For these reasons, a decision was taken to treat data from all crossover RCTs as if they had been derived from parallel trials. However, in recognition of the imperfect nature of this approach, the inclusion of crossover trials in a synthesis was a criterion of downgrading for risk of bias to reflect increased uncertainty in the assembled evidence (see section 8).

4 Dosage

Studies that appeared to employ different dosages for those with abnormal creatinine clearance levels (defined by each study) have been categorised as ‘flexible’ dosing, even if those with normal levels had a fixed dosage. This is to capture that not all patients have received the same dosage.

5 Approach to intention-to-treat analysis

The evidence tables record whether or not studies stated that they performed intention-to-treat (ITT) analysis. The included studies were inconsistent in the use of ITT analysis.

As a result, the reviewers attempted to perform ITT analyses and all patients randomised were included in the denominators for all dichotomous outcomes. Occasionally, the proportion of number of patients which achieved a particular outcome was only reported as a percentage and the total number of patients included in this outcome was not reported; for these studies, the total number of patients achieving the outcome was estimated based on the denominator reported for other dichotomous outcomes or on those randomised. It was not possible to perform an intention-to-treat analysis for continuous outcomes without access to individual patient data.

6 Approach to extracting continuous outcomes

6.1 Creating a normalised outcome from different measuring tools

In order to include as many studies as possible in each individual synthesis, and as the studies reported various tools for measuring pain, the GDG agreed that it was appropriate to convert these different measures onto a common 10-point scale. The GDG felt it was inappropriate to normalise 4-point or smaller scales because of concern about the precision of these scales, and 'converting' them to a normalised 10-point scale may inappropriately inflate the effect estimates of the treatments.

For studies reporting more than 1 tool measuring pain or sleep, the hierarchy of measures preferred when producing single normalised measures was:

- Numerical rating scale
- Visual analogue scale
- Brief Pain Inventory average score or Short-Form McGill for normalised pain measures; Brief Pain Inventory of sleep for normalised sleep interference
- Anything else (for pain, this included Present Pain Intensity for McGill Pain Questionnaire and the Steinbach Pain Inventory)

The above order was based partially from IMMPACT recommendations about tools (Dworkin et al. 2005) and also on data availability (that is, numerical ratings scales were most frequently reported).

6.1.1 Different types of pain scores

Pain was measured in different ways in the included studies. Some studies reported measures such as worst pain, least pain, or pain at the present time, in addition to average pain. Only average mean pain scores were extracted as other measures of pain such as 'worst pain' are difficult to interpret.

Some studies also reported different characteristics of pain, such as allodynia and dysesthetic pain as continuous measures in addition to overall pain. While it was considered important to measure different characteristics of pain to distinguish neuropathic pain from other types of pain that patients may also experience (many patients often have both neuropathic and non-neuropathic pain), this was not consistently reported in the literature. As a result, the ability to meaningfully synthesise these types of pain scores was limited.

6.1.2 Baseline continuous scores

Mean difference from baseline to follow-up was the point of synthesis for all continuous measures. However, mean difference was not always reported in the included studies and it was sometimes necessary to calculate the mean difference from the baseline and follow-up scores.

The mean difference is simply the follow-up measurement minus the baseline value. However, the standard deviation (SD) of mean differences is also required for syntheses. To estimate this, it is necessary to specify the correlation between measurements at the 2 junctures. These were estimated from studies in the effectiveness evidence base. Where a study reports SD at baseline (σ^b), SD at follow-up (σ^f) and the SD of changes between baseline and follow-up (σ^c), the correlation (*C*) between baseline and follow-up for that study may be estimated by:

$$C = \frac{\sigma_b^2 + \sigma_f^2 - \sigma_c^2}{2 \times \sigma_b \times \sigma_f} \quad . \qquad (1)$$

C was calculated for each arm (regardless of treatment assignment) in each study reporting the necessary information. These values were combined by a weighted average according to the number of people in the arm, and the resulting average *C* used to impute SDs of mean differences in studies not reporting them, using the formula:

$$\sigma_c = \sqrt{\sigma_b^2 + \sigma_f^2 - \left(2 \times C \times \sigma_b \times \sigma_f\right)} \quad . \qquad (2)$$

For all calculations, where a baseline pain score was only reported for all patients combined (not for each arm individually), the overall baseline data for all patients in the studies was used for each arm. However, using the same baseline pain score for each arm assumes that randomisation has been adequate and that patients in each group at baseline were similar. Where this is not the case, using this approach could bias results. However, as using the overall baseline pain score allowed mean differences to be calculated and include more studies in the synthesis, the advantages were considered to outweigh the limitations.

7 Synthesis methods

Where possible, meta-analyses were conducted to combine the results of studies for each outcome.

Pairwise meta-analyses were performed using a frequentist approach in Excel.

Network meta-analyses (NMAs) were conducted to simultaneously compare multiple treatments in a single meta-analysis, preserving the randomisation of the randomised controlled trials included in the reviews. This allows all evidence to be combined in a single internally consistent model. A mixed/multiple treatment comparison (MTC) combines both direct and indirect evidence to reduce uncertainty where there are few head-to-head trials, and also provides coherence in the effect estimate producing a more robust estimate of effect. These were used when there were data available on more than two interventions. When there were data available on only two interventions which were not connected by head-to-head evidence, a simple type of network meta-analysis, an indirect treatment comparison (ITC), was used to provide an indirect estimate of the treatment effect between both interventions.

A hierarchical Bayesian NMA was performed using the software WinBUGS version 1.4.3. The models were based on the approach and code provided in the NICE Decision Support Unit's Technical Support Documents on evidence synthesis, particularly Technical Support Document 2 ('A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials'; see http://www.nicedsu.org.uk/).

7.1 Choice of model (random- versus fixed-effects)

A random-effects model was chosen for all meta-analyses because of the relatively heterogeneous populations across the trials. An assumption was made that the different effects are estimating a common distribution. Although a fixed-effects model may have represented a minority of datasets adequately, the value of using a consistent model to estimate all outcomes was felt to outweigh the disadvantage of presenting some models with a superfluous random effects term.

An exception to this principle was in instances where there was only one study for each link in the network. In this case, the assumption was deemed inappropriate and a fixed-effects model was used.

7.2 Dichotomous

In datasets containing studies with 'zero cells' (that is, trials in which no events occurred in 1 or more arm), substantial instability was encountered when performing syntheses. To address this problem, a constant of 0.5 was added to all cell counts (effectively adding 0.5 to the numerator and 1 to the denominator of the proportion). Studies in which all numerators were 0 – that is, studies in which the event of interest was not observed in any relevant arm – were excluded from syntheses, as they do not provide any evidence of effect.

Two alternative models were explored for synthesising dichotomous outcomes: the first relied on a logit link function and produced results in the form of log odds ratios; the second had a complementary log–log ('cloglog') link function, which takes into consideration time to the event, and produces effect estimates in the form of hazard ratios.

There were negligible differences between results from the two types of model. However, it was observed that the cloglog model can be unstable when there are no or few events in either arm (even when a constant was added to studies with zero cells); this problem was particularly common for individual adverse effects. For this reason, logit models were used in the final syntheses (it was also noted that producing results as odds ratios may be more helpful for model validation, as they provide a straightforward point of comparison with frequentist syntheses of direct evidence).

There were 2 exceptions to this principle, where cloglog models were used in preference: withdrawal due to adverse effects (where it might be particularly important to account for differences in follow-up), and the individual adverse effect data that were used in the health economic model (see appendix K). In the latter case, a cloglog model was preferred because it provided results in a form that was more convenient for the model.

7.3 Continuous outcomes

Identity-link models, which rely on a normal likelihood, were used for continuous outcomes. It should be emphasised that these models do not assume that the measures being synthesised are, themselves, normally distributed (which is unlikely to be the case with pain scores); rather, they assume that the sample means are normally distributed (which, according to Central Limit Theorem, will invariably be the case regardless of skewness in the underlying data).

We were unable to include studies that reported continuous data in the form of median differences or as percentage change from baseline in syntheses.

7.4 Prior distributions

Non-informative prior distributions were used in all models. Trial baselines and treatment effects were assigned $N(0, 1000)$ priors, and the between-trial standard deviations used in random-effects models were given $U(0, 5)$ priors. It was felt that this standard deviation was appropriate for both dichotomous and continuous measures. This is recommended in NICE Decision Support Unit's Technical Support Documents (see NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials) for dichotomous outcomes. It was felt appropriate to also use the same value for continuous outcomes since it is not plausible that the SD would be greater than 5 where continuous outcomes are on an 11-point scale.

7.5 Running the model

In the first instance, models were run with 10,000 burn-ins and 50,000 iterations. Three separate chains with different initial values were used. If models did not appear to converge well, they were re-run with more burn-ins.

Syntheses were assessed for any points that significantly deviated from the other data-points and the reasons for any deviate points were investigated.

7.6 Outputs of network meta-analyses

As network meta-analyses do not result in a single point estimate, the results of the meta-analyses were presented in a number of ways.

- Relative effectiveness matrix, showing an estimate of effect for each treatment compared with each of its comparators; an estimate of effect based on direct evidence only (pairwise random-effects meta-analysis) is also presented for comparisons where data are available
- Caterpillar plot of the relative effectiveness of each drug compared with placebo (this includes any direct estimate and also the results of the NMA)
- Probability of each treatment being best
- Median rank with 95% credible interval
- Histograms demonstrating the probability of each treatment at each possible rank ('rankograms').

7.7 Assessing how well the model fit the data

The residual deviance was used and compared to the number of data-points to assess how well the model fit the data. This was recorded in the results section for each model run.

Estimates of residual deviance from the NMA models were also compared with analogous estimates from an 'inconsistency' model, to highlight any inconsistency between direct and indirect evidence (see NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials). There was negligible evidence of inconsistency of this type in any of the syntheses (this is to be expected, as the evidence base is dominated by placebo-controlled trials, with very little head-to-head data). Accordingly, these additional results have not been presented.

8 Quality assessment

GRADE was used to assess the quality of outcomes as specified in the Guidelines Manual.

8.1 GRADE for pairwise meta-analyses

Risk of bias was assessed for:

- Appropriateness of randomisation method
- Adequacy of concealment methods
- Study design – outcomes were downgraded if 50% or more studies in the synthesis were crossover studies
- Comparability of groups at baseline including use of concomitant pain medications, which could have a significant effect on the outcome reported, and which may be incorrectly attributed to the study drug. As a result, individual studies where concomitant drug use between groups (where permitted) was not reported, were treated with caution. Outcomes were downgraded if 50% or more studies that had differing or unclear concomitant drug use at baseline[1][2].
- If the same care was received by each group during the study (this included whether patient in the trial were allowed to vary any concomitant pain medications during the trial)
- Blinding
- Adequacy of length of follow-up (the GDG felt that a minimum of 8 weeks was required for most drugs but that a minimum of 4 weeks was required for topical capsaicin)
- Comparability of those who completed treatment in each group.

Imprecision was assessed as follows:

- Dichotomous outcomes – use optimal information size (OIS) as calculated from http://www.stat.ubc.ca/~rollin/stats/ssize/b2.html.

[1] It should be noted that a large number of studies excluded many concomitant pain medications but still allowed the use of SSRIs. As SSRIs were in the scope of the guideline as a treatment option for neuropathic pain, we recorded that these studies had concomitant pain medication usage and explained this detail in the evidence table.

[2] Throughout the guideline, 'concomitant pain medication' is referred to as 'concomitant drug' usage (not to be confused with concomitant medication for the underlying cause of neuropathic pain or other comorbidities).

- Continuous outcomes – in the absence of guidance on how to determine OIS for continuous outcomes, an OIS of 400 was used as recommended by GRADE.

- Inconsistency was assessed using I^2. If there was considerable heterogeneity (as defined in the Cochrane Handbook for Systematic Reviews of Interventions) between studies in the meta-analysis, the outcome was downgraded 1 level.

8.2 Modified GRADE for network meta-analyses

The use of GRADE to assess the quality of studies addressing a particular review question for pairwise comparisons of interventions is relatively established. However, the use of GRADE to assess the quality of evidence across a network meta-analysis is still a developing methodology. While most criteria for pairwise meta-analyses still apply, it is important to adapt some of the criteria to take into consideration additional factors, such as how each 'link' or pairwise comparison within the network applies to the others. As a result, the following was used when applying modified GRADE to a network meta-analysis.

Risk of bias

In addition to the usual criteria to assess the risk of bias or 'limitations' of studies for each pairwise analysis within a network, the risk of bias was assessed for each direct comparison and then an assessment was made about how the risk of bias from the direct comparisons would affect the indirect comparisons. Additionally, there was an assessment of treatment effect modifiers and if they differed between links in the network.

For studies with a large proportion of studies in a network, some decision rules were applied with respect to downgrading.

- If 50% or more studies in the network were inadequate or unclear for a particular parameter of quality, the outcome was downgraded by 1 level.
- As with pairwise meta-analyses, studies with differences in concomitant drug use between groups, or if concomitant drug use between groups

(where permitted) was not reported, were treated with caution. Additionally, if there were differences in concomitant drug usage among the studies included in different links across the network, the overall outcome was downgraded.

Inconsistency

Inconsistency was assessed for the heterogeneity of individual pairwise comparisons in the network and also for between direct and indirect comparisons, where both were available (that is, where there were ‘loops’ in the network).

Heterogeneity across studies for each direct pairwise meta-analysis was assessed using I^2. This allowed for the assessment of heterogeneity within the included studies using the following decision rules:

- If there was considerable heterogeneity for 1 link or more in a network, the outcome was downgraded 1 level.
- If there were more than 1 link in the network with considerable, substantial or moderate heterogeneity, consider downgrading 2 levels.

To assess for consistency for each pairwise comparison where both direct and indirect evidence are available, the values of the direct and indirect estimates were compared to see if they were similar. (see also section 7.7 above about assessing consistency)

The overall value of tau was also assessed to compare heterogeneity across the network.

Indirectness

As with pairwise meta-analyses, studies included in a network were assessed for how well they fit the PICO (population, intervention, comparator, outcome) specified in the review protocol.

Imprecision

This was assessed for a number of variables:

- Sufficient head-to-head trials in the network.
- Sufficient number of studies to form the network (if there is a high proportion of 'links' formed with only 1 trial, the outcome was downgraded).
- Overall certainty/uncertainty of the effect estimates (size of credible intervals, including for each drug compared to placebo and also size of credible intervals for the overall rankings within the network)[3].
- For networks, imprecision was considered around both the direct and indirect effect estimates.

When assessing imprecision for pairwise comparisons, or for networks with only 1 trial for all 'links' in the network, the confidence interval around the direct estimate was used (since the results were largely led by a non-informative prior).

9 References

Dias, S., Welton, N.J., Sutton, A.J. & Ades, A.E. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated March 2013; available from http://www.nicedsu.org.uk.

Dias, S., Welton, N.J., Sutton, A.J., Caldwell, D.M., Lu, G. & Ades, A.E. NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials. 2011; last updated April 2012; available from http://www.nicedsu.org.uk.

Dworkin RH, Turk DC, Farrar JT, et al (2005) Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113:9–19.

Elbourne DR, Altman DG, Higgins JPT et al. (2002) Meta-analyses involving cross-over trials: methodological issues. International Journal of Epidemiology 31: 140–49.

[3] As ORs are quite difficult to interpret, a confidence interval was considered 'wide' if it was 4 or greater; an outcome was downgraded for imprecision if 50% or more interventions had wide confidence intervals for the OR when they were compared to placebo.

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.