BMJ 2014;349:g5630 doi: 10.1136/bmj.g5630 (Published 24 September 2014)

RESEARCH METHODS & REPORTING

A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis

Network meta-analysis (NMA), combining direct and indirect comparisons, is increasingly being used to examine the comparative effectiveness of medical interventions. Minimal guidance exists on how to rate the quality of evidence supporting treatment effect estimates obtained from NMA. We present a four-step approach to rate the quality of evidence in each of the direct, indirect, and NMA estimates based on methods developed by the GRADE working group. Using an example of a published NMA, we show that the quality of evidence supporting NMA estimates varies from high to very low across comparisons, and that quality ratings given to a whole network are uninformative and likely to mislead.

Milo A Puhan [1], Holger J Schünemann [2], Mohammad Hassan Murad [3], Tianjing Li [4], Romina Brignardello-Petersen [5], Jasvinder A Singh [6], Alfons G Kessels [7], Gordon H Guyatt [2], for the GRADE Working Group

[1]Epidemiology, Biostatistics and Prevention Institute–Epidemiology, Hirschengraben 84, Zurich 8001, Switzerland; [2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Health Sciences Centre, Hamilton, Ontario L8N 3Z5, Canada; [3]Mayo Clinic–Preventive Medicine, Rochester MN, Minnesota 55905, USA; [4]Johns Hopkins Bloomberg School of Public Health–Epidemiology, Baltimore, Maryland, USA; [5]University of Toronto–Clinical Epidemiology and Health Care Research, Toronto, Ontario, Canada; [6]University of Alabama–Clinical Immunology and Rheumatology, Birmingham, Alabama, USA; [7]University of Maastricht, Maastricht, Netherlands

Network meta-analysis (NMA) that simultaneously addresses the comparative effectiveness and/or safety of multiple interventions through combining direct and indirect estimates of effect is rapidly gaining popularity and influence.[1-6] Although NMA approaches appear attractive,[6-8] application of their results requires understanding the quality of the evidence. By quality of evidence, we mean the degree of confidence or certainty one can place in estimates of treatment effects.

NMA is sufficiently new that terminology differs between authors and continues to evolve. Box 1 presents a glossary of terms used in this article.

Rationale for an approach to rate the quality of evidence from NMA

Recently, several articles have provided guidance regarding identification of the evidence for a NMA,[9] statistical aspects of conducting NMA,[10-17] and critical appraisal and interpretation of published NMA.[18 19] Few of these, however, provide explicit guidance on how to rate the quality of the evidence.[4 20 21]

Reports of NMAs often describe the risk of bias of trials included in a NMA (such as method of randomisation, concealment of random allocation, masking, etc).[22-24] For example, a recent NMA compared the effects of coronary artery bypass grafting, various stents, and medical treatment on mortality, myocardial infarction, and the need for revascularisation among patients with stable coronary artery disease. The authors stated that appropriate methods of concealment of random allocation were reported for 71 trials (71%).[25] Fifty six trials (56%) reported blind adjudication of clinical outcomes, and for 69 trials (69%) data from intention to treat analyses were available. Although such an assessment of risk of bias describes the entire body of evidence (that is, all trials contributing evidence to the NMA), it does not acknowledge that the risk of bias is likely to differ across the comparisons of the network.[1] For example, the risk of bias of studies comparing sirolimus eluting stents versus medical treatment may be considerably less than the risk of bias of studies comparing coronary artery bypass grafting with medical treatment. In addition, risk of bias is only one determinant of quality of evidence. Our confidence in effect estimates will, for instance, also decrease if there are large differences in results from study to study (for example, some studies suggest benefit, but others suggest harm) or if results are imprecise (that is, small numbers of patients and resulting wide confidence intervals, see box 2). Furthermore, the popular approach of treatment

Correspondence to: M A Puhan miloalan.puhan@uzh.ch

Box 1: Glossary of terms (in order they appear in the text)

Ranking—Ordering of treatments according to their relative effectiveness. The first ranked treatment is most likely to be the most effective treatment with respect to a particular outcome compared with the other treatments in the network

Direct estimates—Estimate of effect provided by a head-to-head comparison (such as trials of A versus B when A *v* B is the comparison of interest)

Indirect estimates—Estimate of effect provided by two or more head-to-head comparisons that share a common comparator (such as trials of A *v* C and trials of B *v* C when A *v* B is the comparison of interest)

Network—A collection of trials of alternative interventions for a clinical condition that allow, through direct and indirect comparisons, calculation of the relative effects of all treatment versus placebo or standard care, and versus one another, on a particular outcome (for example, fig 1⇓)

Loops—Two or more head-to-head comparisons that contribute to an indirect estimate. First order loops are those loops that involve only a single additional intervention. For example, if we are interested in A versus B, the direct estimates of A versus C and B versus C constitute a first order loop (see red solid line in fig 2⇓). A second order loop would involve two other interventions (such as A *v* C, C *v* D, and D *v* B; see green and blue dashed lines in fig 2⇓). Higher order loops involve additional interventions

Intransitivity—Differences in study characteristics that may modify treatment effect in the direct comparisons (such as A *v* C and B *v* C) that form the basis for the indirect estimate of effect of the comparison of interest (A *v* B), and thus bias the indirect assessment of A versus B. Factors that may modify treatment effects include differing patient characteristics; differing co-interventions; differing extent to which interventions of interest are optimally administered; differing comparators; and differences in measurement of outcome

Heterogeneity—Differences in estimates of effect across studies that assessed the same comparison

Incoherence—Differences between direct and indirect estimates of effect

rankings (for example, probability that coronary artery bypass grafting is the most effective treatment to lower the risk of mortality) will result in misleading inferences when most evidence is low or very low quality, or when evidence supporting higher ranked treatments (such as coronary artery bypass grafting) is much lower quality than evidence supporting lower ranked treatments (such as drug eluting stents). Patients and clinicians may choose a lower ranked treatment with supporting evidence they can trust over a higher ranked treatment with supporting evidence they cannot trust.

The GRADE Working Group and its approach to rate the quality of evidence

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group began in the year 2000 as an informal collaboration of people with an interest in addressing the shortcomings of present grading systems in health care. The GRADE Working Group has developed a sensible and transparent approach to grading quality of evidence (box 2).[20 26-31] The goal of this approach is to provide ratings for the confidence in the estimates of effect for a specific comparison (such as sirolimus eluting stents *v* medical treatment) for all outcomes of importance to patients (such as all-cause mortality, recurrent angina). If all trials are at low risk of bias; if all included populations, interventions, and outcomes are applicable to practice; if trials show similar estimates of treatment effects; if the effect estimates from meta-analysis are precise (for example, narrow 95% confidence interval); and if suspicion of publication bias is low, we will judge the quality of evidence as high (that is, we can be confident that the true effect lies close to that of the estimate of the effect). If, however, trials are at high risk of bias; show inconsistent estimates of effects across trials; included highly selected patients or used surrogate outcomes; if the estimates of treatment effect are imprecise; or if we have a high suspicion of publication bias, we will judge the evidence as lower quality (that is, the confidence in estimates of treatment effect is only moderate, low or very low, box 2).

In this paper we describe the GRADE Working Group's approach to rating the quality of evidence for specific comparisons included in a NMA. Discussion of an approach to rate the confidence in estimates of effect from NMA began at an international meeting on NMA at Johns Hopkins University (Baltimore MD, USA)[1] and with a face-to-face GRADE Working Group meeting in 2010 (Keystone CO, USA). An iterative process including more face-to-face meetings, electronic conferences, email discussions, and multiple iterations of a draft manuscript followed. The final meeting took place at a GRADE Working Group meeting in 2014 (Barcelona, Spain).

The GRADE four-step approach for rating the quality of treatment effect estimates from NMA

Rating the quality of treatment effect estimates from NMA requires best estimates from direct, indirect, and NMA (combined direct + indirect) evidence, as well as quality ratings for the direct and indirect comparisons. We propose the following four steps to assess the quality of treatment effect estimates from NMA (fig 1⇓):

1. Present direct and indirect treatment estimates for each comparison of the evidence network. The direct estimate of effect is provided by a head-to-head comparison (trials of A *v* B), and the indirect estimate is provided by two or more head-to-head comparisons that share a common comparator (for example, we infer the effects of A *v* B from trials of A *v* C and trials of B *v* C).
2. Rate the quality of each direct and indirect effect estimate.
3. Present the NMA estimate for each comparison of the evidence network.
4. Rate the quality of each NMA effect estimate.

Example used for illustration

We use a recent NMA to illustrate the application of the GRADE approach. This article will not present details of the underlying systematic reviews and statistical aspects of the NMA; these are reported elsewhere.[8] In brief, the NMA included randomised trials that compared drug treatments to prevent fragility fractures in individuals with or at risk of osteoporosis. The target population was postmenopausal women at risk of developing fragility fractures, but a small number of eligible trials enrolled men or women irrespective of risk. The drug treatments included bisphosphonates (alendronate, risedronate, zoledronate, and ibandronate), teriparatide, selective oestrogen receptor modulators (raloxifene), denosumab, and calcium and/or vitamin D.

Here, we present the hip fracture outcome data from 40 trials that included 139 647 participants, of whom 2567 (1.8%) had

Box 2: GRADE approach for rating the quality of estimates of treatment effect

Goal of

Provides a rating for the quality of the estimates of effect for a specific comparison and a specific outcome

Ratings

High quality (⊕⊕⊕⊕)—We are very confident that the true effect lies close to that of the estimate of the effect

Moderate quality (⊕⊕⊕O)—We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

Low quality (⊕⊕OO)—Our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect

Very low quality (⊕OOO)—We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect

Starting point

If randomised trials form the evidence base the quality rating starts with high. If observational studies form the evidence base the quality rating starts low. Lack of randomisation typically leads to a initial rating of low

Down rating

The quality rating may be rated down by –1 (serious concern) or –2 (very serious concern) for the following reasons

- Risk of bias (such as failure to conceal random allocation or blind participants[28] in randomised controlled trials or failure to adequately control for confounding in observational studies)
- Inconsistency (such as heterogeneity of estimates of effects across trials[29])
- Indirectness (such as surrogate outcomes, study populations or interventions that differ from those of interest,[20] or intransitivity;[19] for further explanation see box 3)
- Imprecision (for example, 95% confidence intervals are wide and include or are close to null effect[30])
- Publication bias[31]

Up rating

Rating up is typically applied only to observational studies; the most common reason is for a large or very large effect seen over a short period of time and altering a clear downward trajectory

Explanations

For reasons of transparency and for better understanding of the ratings, reasons for rating are documented, typically in a summary table's footnotes

Box 3: Clarification of GRADE rating down for indirectness

Indirectness, a term established by the GRADE Working Group,[20] refers to two different concepts.

One concept relates to differences between the question of interest and the body of evidence that is identified and used to inform the question. We may rate down the quality for this type of indirectness when patients of interest (as defined by the question) overlap only partly with patients enrolled in trials (for example, the population of interest is the very elderly, few of who participated in the trials); interventions of interest differ from those regimens tested in trials (for example, the intensity of anticoagulation control differed in trials compared with the community setting); or outcomes of interest differ from those measured in trials (for example, trials in diabetes measured blood glucose, a surrogate endpoint, rather than cardiovascular events).

The second concept on indirectness, particularly relevant for NMA, relates to biased evidence from indirect comparisons. In accordance with the NMA literature, in this article we refer to this second concept as intransitivity (see box 1 for definition of intransitivity).[15 16]

a hip fracture. The results presented in this paper are identical to those of the primary report,[8] and we did not perform any new NMA for this paper. We did, however, apply our new GRADE approach to rating the quality of evidence of each comparison (this was not done in the original article). Figure 2⇓ shows the evidence network for the available direct comparisons.

Step 1: Presenting direct and indirect effect estimates and 95% CI

Making valid inferences on the basis of a NMA requires understanding of both the direct and indirect evidence that contributes to the NMA effect estimates. Several approaches exist for calculating indirect estimates.[12 32 33] For the example presented here we use a method referred to as node splitting, which separates evidence on a particular comparison (a "node") into direct and indirect estimates of treatment effect.[12] For example, direct evidence for the comparison of alendronate versus raloxifene in our fracture prevention example shows an odds ratio of 0.49.[8] Because the trial directly comparing the two agents is small, the 95% confidence interval is wide (0.04 to 5.45, fig 1⇓). The indirect evidence (odds ratio 0.53, 95% confidence interval 0.30 to 0.90) includes a first order loop (first order loops are those loops that involve only a single additional intervention, such as vitamin D plus calcium, see red solid line in fig 2⇓) and second order loops (loops that involve two other interventions, such as calcium, vitamin D, and placebo, see green and blue dashed lines in fig 2⇓).

Step 2: Rating of quality of direct and indirect effect estimates

Investigators rate the quality of evidence separately for direct and indirect evidence. The confidence estimates for the direct comparisons involve an application of the GRADE principles (box 2) to each comparison for which head-to-head trials are available. For the network of drugs to prevent osteoporotic fractures, we found seven direct comparisons to warrant high or moderate confidence and nine direct comparisons to warrant low or very low confidence (table 1⇓).

Depending on the size and structure of the evidence network, one, few, or many loops can contribute indirect evidence to the comparisons of interest. To keep the quality rating of the indirect evidence manageable, we suggest a focus on first order loops, which usually contribute most information to the indirect estimate. To identify the relevant loops a network graph such

as figure 2⇓ is needed (red solid line represents the first order loop for the indirect comparison of alendronate *v* raloxifene).

The rating of the quality of the indirect estimate is then based on the ratings of the two pairwise estimates (such as A *v* C and B *v* C) that contribute to the indirect estimate of the comparison of interest (A *v* B); these ratings can follow established GRADE guidance.[26] For example, when comparing alendronate versus raloxifene, the comparisons of alendronate versus vitamin D plus calcium and raloxifene versus vitamin D plus calcium (fig 2⇓, red solid line) create the first order loop. The lower confidence rating of the two direct comparisons constitutes the confidence rating of the indirect comparison. In this case, for both comparisons, the confidence rating is moderate: therefore, the initial rating of the indirect evidence warrants moderate confidence.

There is, however an additional issue that may further reduce confidence in estimates from the indirect comparison: intransitivity (see box 1). If the trials forming the basis for the indirect estimate (such as trials of A *v* C and of B *v* C) differ in important ways the likelihood of intransitivity may be high. As a consequence the indirect estimate of the comparison of interest (A *v* B) may be biased. In the presence of intransitivity we would rate down further from the lower of the confidence ratings of the contributing direct comparisons.

Consider, for example, the indirect comparison for risedronate versus vitamin D plus calcium (fig 2⇓). The trials with placebo as common comparator provide most of the indirect evidence. Risedronate was tested in 20 trials for the prevention of fragility fractures. In half of these trials, patients were using glucocorticoid treatment or had a chronic disease that might modify bone metabolism (such as inflammatory bowel disease).[8] This contrasts with the trials of vitamin D plus calcium versus placebo, in which participants were included only if they did not take drugs and did not have diseases that modify bone metabolism.[34] As a consequence of these differences between the trials of risedronate and vitamin D plus calcium versus the common comparator placebo, we decided to down rate the indirect comparison of risedronate versus vitamin D plus calcium for intransitivity.

It is conceivable that a substantial proportion of indirect comparisons of any NMA warrant down rating for indirectness because of these two reasons. Although we suggest a low threshold for down rating for indirectness, authors should be explicit and report reasons for down rating in the footnotes of the table that presents the direct, indirect, and network estimates of effect. For the network of drugs to prevent osteoporotic fractures, we found 10 indirect comparisons to be of high or moderate quality, respectively, and 41 indirect comparisons to be of low or very low quality, respectively (table 1⇓).

Steps 3 and 4: Presenting and rating of quality of NMA effect estimates

If only direct or indirect evidence is available for a given comparison, the network quality rating will be based on that estimate. When, for a particular comparison, both direct and indirect evidence are available, we suggest using the higher of the two quality ratings as the quality rating for the NMA estimate (for example, moderate quality if quality of the direct estimate is moderate and quality of the indirect estimate is low).

There are two reasons we have chosen this approach. First, if direct and indirect estimates are similar (coherent, see box 1), the lower quality estimate can only bolster the higher (it would make no sense to add evidence that would lower the quality of estimates). Second, in general, we expect the higher rated estimate to be the more precise (and thus dominating) body of evidence.

In the rarer instances in which the less precise estimate warrants higher confidence it likely means that there are no other reasons for down rating that estimate. On the other hand, if the more precise estimate warrants lower confidence than the less precise estimate, there must be serious problems (risk of bias, inconsistency, publication bias, indirectness). If direct and indirect are coherent, the serious problems with risk of bias in the lower confidence are unlikely to have biased the results. If there is serious incoherence then we default to the following guidance regarding what to do in the presence of incoherence.

The assessment of coherence (others use different terminology such as inconsistency) addresses the assumption that direct and indirect evidence are similar enough to be pooled. A commonly used approach to investigate coherence is to test the statistical significance of the difference between direct and indirect estimates.[11 12 29] In addition, the magnitude of differences between the direct and indirect estimates should bear on addressing incoherence.

Consider table 2⇓, which presents results from a NMA of the impact of alternative surgical approaches to open tibial fractures on reoperation (from Foote CJ, Guyatt GH, Vignesh KN, et al "Systematic review of prospective investigation of surgical treatment of open tibial shaft fractures (SPRINT review): a network meta-analysis" submitted for publication). In the comparison of unreamed versus reamed nailing, the direct estimate suggests unreamed is superior. The indirect evidence also suggests unreamed is superior, but the effect is much larger, the confidence intervals of the two estimates are virtually non-overlapping, and the statistical test of interaction generates a P value of 0.02. This suggests major incoherence between direct and indirect estimates. On the other hand, for the comparison of unreamed nailing versus external fixation (table 2⇓) we would conclude the results are coherent.

In the face of large incoherence in a particular comparison we do not advocate discarding or modifying the NMA (for instance, by excluding the incoherent data) without a strong rationale. NMA authors can guide users of the NMA in one of two ways. The first is to focus attention on the direct or indirect estimate warranting greater confidence, rather than the NMA estimate, as the best estimate of effect. This is the approach authors used in the NMA of open tibial fractures. An alternative is to focus on the NMA estimate but rate down the quality of that estimate for incoherence (in this example, also for imprecision, thus leading to a judgment of low quality).

The optimal strategy is likely to depend on the circumstances. If the difference in quality between the two estimates is large, and one of the two is of higher quality, the former approach may be desirable. If the difference in quality between the two estimates is smaller, and neither is of high quality, using the NMA estimate and rating down for incoherence may be preferable. When there is only indirect evidence it is not possible to assess incoherence.[33] In such situations issues regarding intransitivity may warrant particular attention, and the threshold for rating down for intransitivity may be lower.

In the example of preventing osteoporotic fractures, table 1⇓ presents the NMA estimates and the final quality ratings. For most of the comparisons, there is only indirect evidence (such as alendronate *v* zoledronate), and the quality rating of the indirect comparison also represents the quality of the NMA estimate. For the comparison of vitamin D plus calcium versus risedronate, direct evidence had very low confidence rating and contributed substantially more to the NMA estimate than indirect

evidence; therefore, the quality rating for the NMA estimate was also very low. Across the network, we found three comparisons (5% of all comparisons) of high quality, 13 (24%) of moderate quality, 19 (35%) of low quality, and 20 (36%) of very low quality.

Finally, one further criterion warrants consideration. While estimates from both direct and indirect may cross a threshold that warrants rating down for imprecision, the pooled network estimate, because it is more precise, may not. For example, we rated down both the direct and indirect comparisons of calcium versus calcium plus vitamin D for imprecision (table 1⇓). Since the pooled estimate was more precise, therefore increasing our confidence that calcium plus vitamin D is more effective than calcium alone, we did not rate down the NMA estimate for imprecision.

Variability of quality of treatment effect estimates and ranking

Quality of estimates can vary greatly across comparisons within the network. Indeed, in our illustrative example, quality varied from high to very low (table 1⇓). In making inferences regarding choice of intervention, recognising the quality of each comparison is far more valuable than the single risk of bias assessment across an evidence network typically reported in most NMA articles.[22-25]

An example of the necessity for rating the quality of individual paired comparisons arises from the initial report of the fracture NMA we present here. Using the standard ranking approach in NMA, the authors concluded teriparatide had the largest fracture reduction of the 10 treatments studied (odds ratio 0.42 against no treatment, table 1⇓) and the highest probability of being ranked first across the treatments. Our quality ratings of teriparatide against placebo and other comparators are, however, low or very low (table 1⇓). Other agents (zoledronate or denosumab) had high or moderate confidence ratings of superiority over placebo and over vitamin D plus calcium. The quality ratings suggest that clinicians and patients seeking a drug that prevents hip fractures will be better off choosing zoledronate or denosumab than teriparatide.

What the GRADE guidance adds to existing NMA guidance documents

A wealth of literature addressing NMA has accumulated over recent years. For example, Cipriani and colleagues provided excellent guidance on key statistical aspects of NMA.[10] A task force of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) published three documents addressing the conduct and interpretation of NMA as well as a checklist for critical appraisal of NMA.[18 19 35] A recent overview of reporting practices paves the way for an extension of the PRISMA statement addressing NMA.[36]

Some of these documents addressed risk of bias for the entire network or specific comparisons of a NMA.[14 18 19 35] Only one, a more statistically oriented paper, provides guidance on how to rate the quality of indirect and NMA estimates considering not only risk of bias but also other criteria that affect confidence in estimates of effect (box 2 and criteria specific to NMA).[21] The four-step approach of the GRADE Working Group fills a gap by providing guidance to determine quality ratings for each estimate of effect in a NMA.

Research needs

There are a number of studies that would be useful to refine the four-step approach presented here. A previous study showed that inter-rater agreement is high if the raters are familiar with GRADE and if calibration exercises are done.[37] It is important to conduct inter-rater agreement studies for NMA in order to identify those aspects of the rating process that require additional guidance and calibration exercises. Meta-epidemiological studies addressing the effects of specific criteria (such as intransitivity) on estimates of effect provided by NMA would be useful to inform how readily one should down rate the quality.

Finally, we do not currently support the use of weights (reflecting the amount of information) to decide if the quality rating of the direct or indirect estimate should determine the quality of the NMA estimate. Statistical approaches to determine weights are already incorporated in standard statistical packages. There is, however, little experience in the interpretation and use of such weights in different NMA (that may differ by, for example, their geometry). Studies that inform the optimal use of weights may lead to a revised approach for generating quality ratings for NMA estimates.

Conclusion

The GRADE Working Group approach following four steps highlights the necessity for authors of NMA to present direct, indirect, and NMA estimates as well as quality ratings for all direct comparisons. If authors do not present these estimates, scepticism regarding any inferences from the NMA is warranted.

Contributors: MP had the study idea, developed the grading approach, drafted the article and is guarantor of the article; HJS had the study idea and contributed to development of the grading approach and critical revision of the article; MHM and TL contributed to development of the grading approach, data analyses and critical revision of the article; RBP, JAS and AGK contributed to development of the grading approach and critical revision of the article; GG had the study idea, developed the grading approach and drafted the article.

Funding: No funding support.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and have no relevant interests to declare.

1 Li T, Puhan MA, Vedula SS, Singh S, Dickersin K; Ad Hoc Network Meta-analysis Methods Meeting Working Group. Network meta-analysis-highly attractive but more methodological research is needed. *BMC Med* 2011;9:79.
2 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;23:3105–24.
3 Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002;21:2313–24.
4 Mills EJ, Ioannidis JP, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA* 2012;308:1246–53.
5 Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* 2009;338:b1147.
6 Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;373:746–58.
7 Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. *BMC Med* 2009;7:2.
8 Murad MH, Drake MT, Mullan RJ, Mauck KF, Stuart LM, Lane MA, et al. Clinical review. Comparative effectiveness of drug treatments to prevent fragility fractures: a systematic review and network meta-analysis. *J Clin Endocrinol Metab* 2012;97:1871–80.
9 Hawkins N, Scott DA, Woods B. How far do you go? Efficient searching for indirect evidence. *Med Decis Making* 2009;29:273–81.
10 Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med* 2013;159:130–7.
11 Veroniki AA, Vasiliadis HS, Higgins JPT, Salanti G. Evaluation of inconsistency in networks of interventions. *Int J Epidemiol* 2013;42:332–45.
12 Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* 2010;29:932–44.

13 Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis. *Med Decis Making* 2013;33:671–8.
14 Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One* 2013;8:e76654.
15 Salanti G, Higgins JPT, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008;17:279–301.
16 Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med* 2013;11:159.
17 Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. *J Clin Epidemiol* 2014;67:672–80.
18 Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014;17:157–73.
19 Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. *Value Health* 2011;14:417–28.
20 Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol* 2011;64:1303–10.
21 Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. *PLoS ONE* 2014;9:e99682.
22 Stettler C, Wandel S, Allemann S, Kastrati A, Morice MC, Schömig A, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. *Lancet* 2007;370:937–48.
23 Palmerini T, Biondi-Zoccai G, Riva DD, Mariani A, Savini C, Di Eusanio M, et al. Risk of stroke with percutaneous coronary intervention compared with on-pump and off-pump coronary artery bypass graft surgery: Evidence from a comprehensive network meta-analysis. *Am Heart J* 2013;165:910–7.e14.
24 Pandor A, Gomersall T, Stevens JW, Wang J, Al-Mohammad A, Bakhai A, et al. Remote monitoring after recent hospital discharge in patients with heart failure: a systematic review and network meta-analysis. *Heart* 2013;99:1717-26.
25 Windecker S, Stortecky S, Stefanini GG, da Costa BR, Rutjes AW, Di Nisio M, et al. Revascularisation versus medical treatment in patients with stable coronary artery disease: network meta-analysis. *BMJ* 2014;348:g3859.
26 Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.
27 Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;64:380–2.
28 Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15.
29 Guyatt GH, Oxman AD, Kunz R, , Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol* 2011;64:1294–302.
30 Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol* 2011;64:1283–93.
31 Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 2011;64:1277–82.
32 Kessels A, ter Riet G, Puhan M, et al. A simple regression model for network meta-analysis. *OA Epidemiology* 2013;1:7.
33 Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* 2012;3:98–110.
34 Avenell A, Mak JCS, O'Connell D. Vitamin D and vitamin D analogues for preventing fractures in post-menopausal women and older men. *Cochrane Database Syst Rev* 2014;4:CD000227.
35 Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. *Value Health* 2011;14:429–37.
36 Hutton B, Salanti G, Chaimani A, Caldwell DM, Schmid C, Thorlund K, et al. The quality of reporting methods and results in network meta-analyses: an overview of reviews and suggestions for improvement. *PLoS One* 2014;9:e92508.
37 Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol* 2013;66:736–42; quiz 742.e1–5.

Accepted: 22 August 2014

Cite this as: *BMJ* 2014;349:g5630

Tables

Table 1| Estimates of effects and quality ratings for comparison of drugs to prevent osteoporotic hip fractures

	Direct evidence		Indirect evidence		Network meta-analysis	
Comparison	Odds ratio (95% confidence interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence
Teriparatide *v* placebo	—	—	0.42 (0.10 to 1.82)	Very low‡,**	0.42 (0.10 to 1.82)	Very low
Denosumab *v* placebo	—	—	0.50 (0.27 to 0.86)	High	0.50 (0.27 to 0.86)	High
Raloxifene *v* placebo	0.84 (0.63 to 1.13)	Moderate‡	0.96 (0.53 to 1.78)	Low‡,¶	0.87 (0.63 to 1.22)	Moderate
Zoledronate *v* placebo	—	—	0.50 (0.33 to 0.74)	High	0.50 (0.34 to 0.73)	High
Risedronate *v* placebo	0.17 (0.05 to 0.59)	Low*,‡‡	0.54 (0.36 to 0.75)	Low**	0.48 (0.31 to 0.66)	Low
Ibandronate *v* placebo	—	—	0.49 (0.21 to 1.20)	Very low‡,**	0.49 (0.21 to 1.20)	Very low
Alendronate *v* placebo	—	—	0.45 (0.27 to 0.68)	Moderate ¶	0.45 (0.27 to 0.68)	Moderate
Vitamin D *v* placebo	1.25 (0.82 to 1.89)	Low*,‡	1.08 (0.61; 1.91)	Low‡,¶	1.13 (0.94 to 1.34)	Low
Vitamin D+calcium *v* placebo	0.83 (0.73 to 0.96)	Moderate*	0.54 (0.29 to 0.94)	Low**	0.81 (0.68 to 0.96)	Moderate
Calcium *v* placebo	—	—	1.14 (0.82 to 1.59)	Moderate¶	1.14 (0.82 to 1.59)	Moderate
Denosumab *v* teriparatide	—	—	1.17 (0.24 to 5.54)	Low‡,¶	1.17 (0.24 to 5.54)	Low
Raloxifene *v* teriparatide	—	—	2.05 (0.47 to 9.47)	Very low‡,**	2.05 (0.47 to 9.47)	Very low
Zoledronate *v* teriparatide	—	—	1.18 (0.26 to 5.30)	Very low‡,¶,‡‡	1.18 (0.26 to 5.30)	Very low
Risedronate *v* teriparatide	—	—	1.12 (0.25 to 4.98)	Very low‡,¶,‡‡	1.12 (0.25 to 4.98)	Very low
Ibandronate *v* teriparatide	—	—	1.11 (0.22 to 6.42)	Very low‡,**	1.11 (0.22 to 6.42)	Very low
Alendronate *v* teriparatide	—	—	1.02 (0.24 to 4.82)	Very low‡,**	1.02 (0.24 to 4.82)	Very low
Vitamin D *v* teriparatide	—	—	2.67 (0.63 to 11.97)	Very low‡,**	2.67 (0.63 to 11.97)	Very low
Vitamin D+calcium *v* teriparatide	2.00 (0.50 to 8.33)	Low*,‡	Not estimable††	Not estimable††	1.92 (0.45 to 8.42)	Low
Calcium *v* teriparatide	—	—	2.69 (0.63 to 12.23	Very low‡,**	2.69 (0.63 to 12.23)	Very low
Raloxifene *v* denosumab	—	—	1.76 (0.95 to 3.41)	Low‡,¶	1.76 (0.95 to 3.41)	Low
Zoledronate *v* denosumab	—	—	1.02 (0.54 to 1.93)	Low‡,‡‡	1.02 (0.54 to 1.93)	Low
Risedronate *v* denosumab	—	—	0.96 (0.50 to 1.78)	Very low‡,**,‡‡	0.96 (0.50 to 1.78)	Very low
Ibandronate *v* denosumab	—	—	0.98 (0.36 to 2.79)	Low‡,¶	0.98 (0.36 to 2.79)	Low
Alendronate *v* denosumab	—	—	0.90 (0.45 to 1.78)	Low‡,¶	0.90 (0.45 to 1.78)	Low
Vitamin D *v* denosumab	—	—	2.28 (1.28 to 4.16)	Moderate¶	2.28 (1.28 to 4.16)	Moderate
Vitamin D+calcium *v* denosumab	1.67 (1.02 to 2.70)‡	Moderate	Not estimable††	Not estimable††	1.64 (0.97 to 2.87)	Moderate
Calcium *v* denosumab	—	—	2.33 (1.25 to 4.40)	Moderate¶	2.33 (1.25 to 4.40)	Moderate
Zoledronate *v* raloxifene	—	—	0.57 (0.35 to 0.93)	Low¶,‡‡	0.57 (0.35 to 0.93)	Low
Risedronate *v* raloxifene	—	—	0.55 (0.31 to 0.84)	Very low**,‡‡	0.55 (0.31 to 0.84)	Very low
Ibandronate *v* raloxifene	—	—	0.55 (0.23 to 1.42)	Very low‡,**	0.55 (0.23 to 1.42)	Very low
Alendronate *v* raloxifene	0.49 (0.04 to 5.45)	Low§	0.53 (0.30 to 0.90)	Moderate¶	0.51 (0.29 to 0.87)	Moderate
Vitamin D *v* raloxifene	—	—	1.30 (0.89 to 1.86)	Low**	1.30 (0.89 to 1.86)	Low
Vitamin D+calcium *v* raloxifene	0.88 (0.51 to 1.54)	Moderate‡	0.96 (0.63 to 1.49)	Low¶,‡‡	0.94 (0.66 to 1.31)	Moderate
Calcium *v* raloxifene	—	—	1.31 (0.83 to 2.06)	Very low‡,**	1.31 (0.83 to 2.06)	Very low
Risedronate *v* zoledronate	—	—	0.96 (0.56 to 1.49)	Low‡‡	0.96 (0.56 to 1.49)	Low
Ibandronate *v* zoledronate	—	—	0.97 (0.39 to 2.55)	Very low‡,¶,‡‡	0.97 (0.39 to 2.55)	Very low
Alendronate *v* zoledronate	—	—	0.90 (0.51 to 1.51)	Low‡,‡‡	0.90 (0.51 to 1.51)	Low

Table 1 (continued)

Comparison	Direct evidence		Indirect evidence		Network meta-analysis	
	Odds ratio (95% confidence interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence
Vitamin D *v* zoledronate	—	—	2.26 (1.50 to 3.42)	Low¶,‡‡	2.26 (1.50 to 3.42)	Low
Vitamin D+calcium *v* zoledronate	1.64 (1.16 to 2.17)	High	Not estimable††	Not estimable††	1.63 (1.16 to 2.30)	High
Calcium *v* zoledronate	—	—	2.29 (1.44 to 3.66)	Low¶,‡‡	2.29 (1.44 to 3.66)	Low
Ibandronate *v* risedronate	—	—	1.02 (0.43 to 2.66)	Very low‡,**,‡‡	1.02 (0.43 to 2.66)	Very low
Alendronate *v* risedronate	—	—	0.93 (0.54 to 1.62)	Very low**,‡‡	0.93 (0.54 to 1.62)	Very low
Vitamin D *v* risedronate	—	—	2.35 (1.63 to 3.76)	Very low**,‡‡	2.35 (1.63 to 3.76)	Very low
Vitamin D+calcium *v* risedronate	1.92 (0.84 to 4.35)	Very low*,†,‡	5.88 (1.79 to 25.00)	Low¶,‡‡	1.69 (1.27 to 2.54)	Low
Calcium *v* risedronate	—	—	2.39 (1.56 to 4.04)	Very low**,‡‡	2.39 (1.56 to 4.04)	Very low
Alendronate *v* ibandronate	—	—	0.92 (0.34 to 2.32)	Low‡,¶	0.92 (0.34 to 2.32)	Low
Vitamin D *v* ibandronate	—	—	2.32 (0.92 to 5.54)	Very low‡,**	2.32 (0.92 to 5.54)	Very low
Vitamin D+calcium *v* ibandronate	1.72 (0.76 to 3.85)	Low‡	Not estimable††	Not estimable††	1.69 (0.69 to 3.84)	Low
Calcium *v* ibandronate	—	—	2.36 (0.92 to 5.87)	Very low‡,**	2.36 (0.92 to 5.87)	Very low
Vitamin D *v* alendronate	3.70 (1.20 to 11.11)	Moderate*	2.38 (1.49 to 3.85)	Moderate¶	2.54 (1.63 to 4.16)	Moderate
Vitamin D+calcium *v* alendronate	1.59 (1.03 to 2.44)	Moderate*	2.78 (1.14 to 8.33)	Moderate¶	1.82 (1.24 to 2.90)	Moderate
Calcium *v* alendronate	4.55 (0.47 to 50.00)	Very low*,§	2.56 (1.54 to 4.35)	Moderate¶	2.56 (1.57 to 4.34)	Moderate
Vitamin D+calcium *v* vitamin D	1.03 (0.68 to 1.54)	Low*,‡	0.65 (0.48 to 0.85)	Low¶,‡‡	0.72 (0.57 to 0.91)	Low
Calcium *v* vitamin D	—	—	1.01 (0.72 to 1.44)	Low**	1.01 (0.72 to 1.44)	Low
Calcium *v* calcium+vitamin D	1.21 (0.89 to 1.66)	Low*,‡	3.43 (0.26 to 160.40)	Very low‡,**,‡‡	1.40 (1.03 to 1.95)	Moderate§§

*Limitations (risk of bias). †Inconsistency. ‡Imprecision. §Severe imprecision. ¶Contributing direct evidence of moderate quality. **Contributing direct evidence of low or very low quality. ††Cannot be estimated because the drug was not connected in a loop in the evidence network. ‡‡Indirectness because of questionable comparability of trial populations to target population of NMA (postmenopausal women) or because of intransitivity. §§Greater precision.

Table 2| Illustration of coherence and incoherence from a network meta-analysis of alternative surgical approaches to open tibial fractures

	Direct evidence		Indirect evidence		Network meta-analysis	
Comparison	Odds ratio (95% confidence interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence	Odds ratio (95% credible interval)	Quality of evidence
Unreamed *v* reamed	0.74 (0.45 to 1.24)	⊕⊕⊕O* Moderate	0.07 (0.01 to 0.46)	⊕⊕OO†,‡ Low	0.62 (0.37 to 1.03)	⊕⊕OO§ Low
Unreamed *v* external fixation	0.39 (0.23 to 0.65)	⊕⊕⊕O¶ Moderate	0.35 (0.08 to 1.56)	⊕⊕OO*,† Low	0.38 (0.23 to 0.62)	⊕⊕⊕O Moderate

*Imprecision. †Contributing direct evidence of moderate quality. ‡Indirectness. §Incoherence. ¶Limitations (risk of bias).

Figures

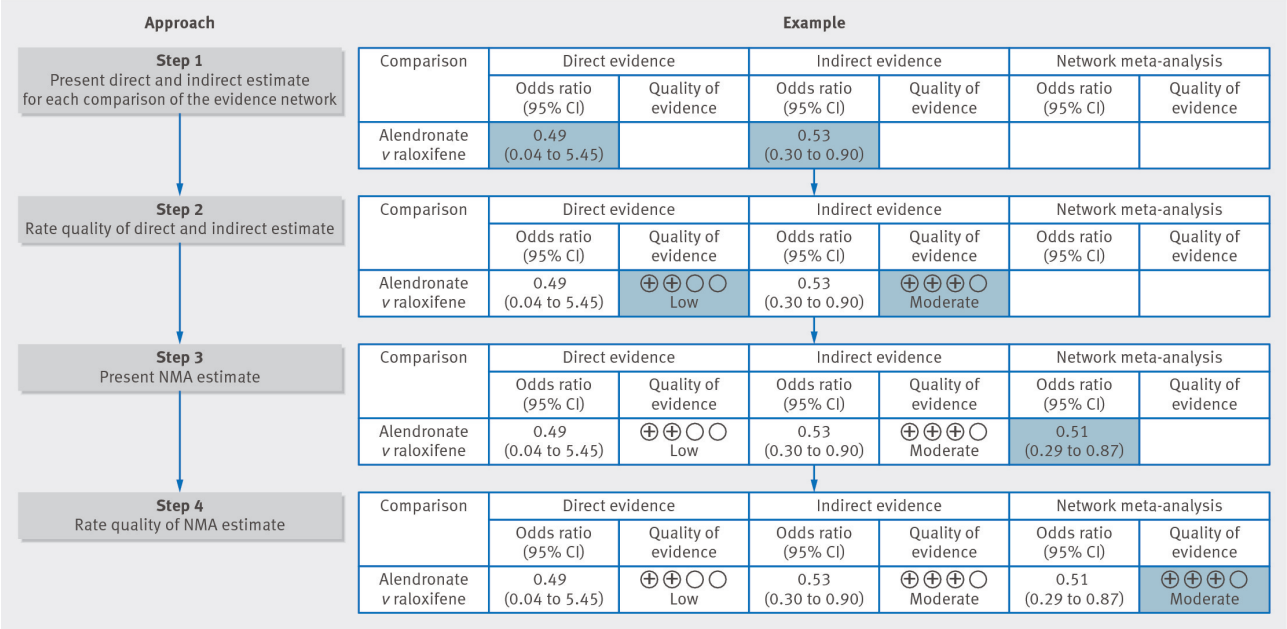

Fig 1 Approach for rating the quality of network meta-analysis (NMA) estimates

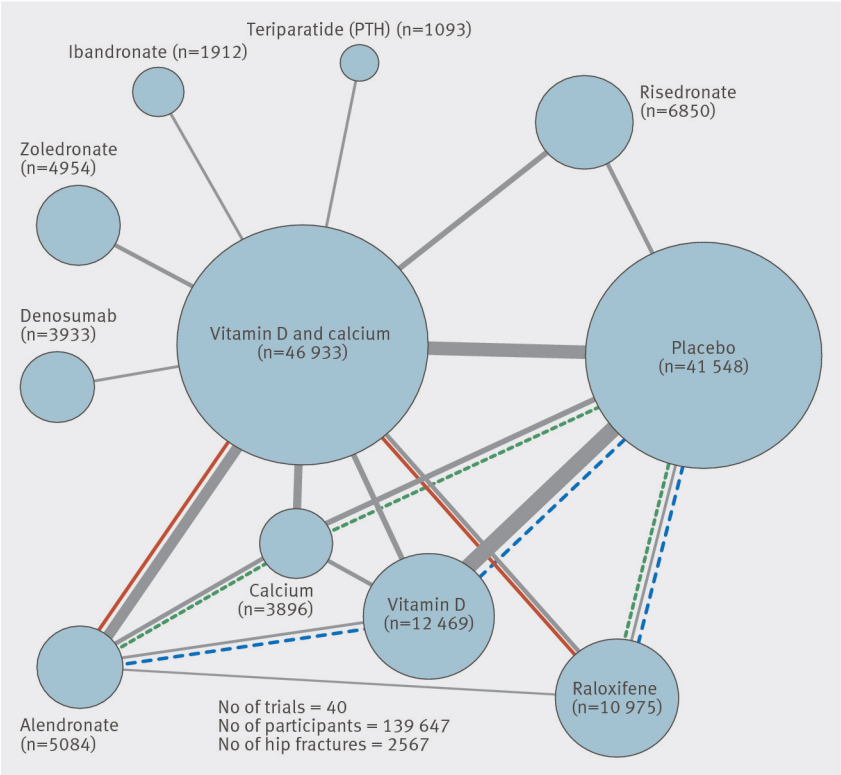

Fig 2 Evidence network of randomised trials comparing the effects of drugs to prevent osteoporotic hip fractures. The size of the circle is proportional to the number of participants randomised to that treatment. Width of the lines is proportional to the number of trials for that comparison. Coloured dashed lines refer to loops for indirect evidence (see text).