

Network Meta-analysis: Users' Guide for Surgeons: Part II – Certainty

Harman Chaudhry MD, Clary J. Foote MD, Gordon Guyatt MD, MSc,
Lehana Thabane PhD, Toshi A. Furukawa MD, PhD, Brad Petrisor MD,
Mohit Bhandari MD, PhD

Received: 4 December 2014 / Accepted: 26 March 2015 / Published online: 14 April 2015
© The Association of Bone and Joint Surgeons® 2015

Abstract In the previous article (Network Meta-analysis: Users' Guide for Surgeons—Part I, Credibility), we presented an approach to evaluating the credibility or methodologic rigor of network meta-analyses (NMA), an innovative approach to simultaneously addressing the relative effectiveness of three or more treatment options for a given medical condition or disease state. In the second part of the Users' Guide for Surgeons, we discuss and demonstrate the application of criteria for determining the certainty in effect sizes and directions associated with a given treatment option through an example pertinent to clinical orthopaedics.

Introduction

A network meta-analysis (NMA) is a method of conducting multiple treatment comparisons using direct evidence (as reported in head-to-head studies, A versus B) and indirect evidence (making inferences about A vs C through a common comparator B) to determine the relative merits of alternative management strategies. As discussed in Part I of the Users' Guide for Surgeons, determining the credibility of the NMA methodology is an important first step in assessing how trustworthy reported treatment effects are. Even the most robust methodology cannot overcome serious limitations in the available evidence. Therefore, the second judgment, assessing certainty in estimates in effect sizes and directions, is crucial. Certainty also may be referred to as the quality of evidence or confidence in effects; however, use of the term “certainty” prevents confusion with quality of individual trials (risk of bias) or confidence intervals (CIs), respectively.

In general, certainty in effect sizes and directions varies widely among comparisons in a network [20, 21]. A NMA therefore can provide results warranting moderate or high certainty for some comparisons and low or very low certainty for others. For this reason, certainty in the estimates of effects requires a detailed evaluation of each pairwise treatment comparison conducted in a NMA.

A recent publication presented detailed criteria from the Grades of Recommendation, Assessment, Development and Evaluation (GRADE) Working Group for determining certainty in estimates from a NMA [23]. Newer NMAs likely will incorporate these guidelines for readers; however, in the absence of such reporting, authors should provide information addressing key questions that are related to certainty (Table 1) [9, 10].

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

H. Chaudhry (✉), C. J. Foote, G. Guyatt, B. Petrisor,
M. Bhandari
Division of Orthopaedic Surgery, McMaster University, 293
Wellington Street N, Suite 110, Hamilton, ON L8L 2X2, Canada
e-mail: harman.chaudhry@medportal.ca

H. Chaudhry, C. J. Foote, G. Guyatt, L. Thabane, M. Bhandari
Department of Clinical Epidemiology and Biostatistics,
McMaster University, Hamilton, ON, Canada

T. A. Furukawa
Department of Health Promotion and Human Behavior, Kyoto
University Graduate School of Medicine/School of Public
Health, Kyoto, Japan

T. A. Furukawa
Department of Clinical Epidemiology, Kyoto University
Graduate School of Medicine/School of Public Health, Kyoto,
Japan

Table 1. Guide for determining certainty in the results

Is there low risk of bias for each comparison?
Are there concerns about publication bias?
Are the results of individual studies in direct comparisons consistent?
Do indirect comparisons respect the transitivity principle?
Are the results of direct and indirect comparisons consistent?
Are confidence or credible intervals sufficiently narrow?
Are the treatment rankings trustworthy?

Evaluating Certainty in Estimates of Effect

Is There Low Risk of Bias for Each Comparison?

As is true with a conventional meta-analysis, conclusions of a NMA ultimately are limited by the risk of bias of the included studies. Studies that are compromised by high risk of bias are likely to result in misleading estimates from the NMA, regardless of how well the authors of the NMA have collected and summarized the evidence.

Because they are limited in controlling imbalance resulting from differences in age, sex, illness severity, and other factors that influence patient outcome (prognostic factors), observational studies are at higher risk of bias than are randomized controlled trials (RCTs). Although observational studies may be included in NMAs, this should be done with careful consideration, and the limitations of doing so should be duly acknowledged. RCTs, through randomization and large sample size, potentially address issues of prognostic imbalance. However, RCTs vary with respect to risk of bias depending on the rigor with which they: (1) conceal randomization; (2) blind participants, caregivers, data collectors, adjudicators of outcome and data analysts; and (3) achieve complete followup [3, 22]. A NMA should present an assessment of risk of bias using an instrument that systematically evaluates all important elements [2]. A modification of the Cochrane risk of bias tool for RCTs is an example of such an instrument [11].

Risk of bias is likely to differ across paired comparisons (surgical treatment A vs B, A vs C, B vs C, and so on), so that reporting the overall risk of bias across a network (for instance, reporting the number of studies at high and low risk of bias among all those included), is of limited use in interpreting results. More useful is information allowing assessment of the extent of risk of bias for each of the paired comparisons under investigation.

Returning to the NMA of open tibial shaft fracture management by Foote et al. [7] in the case scenario presented in Part I of the Users' Guide for Surgeons, we saw that the NMA had restricted eligibility to RCTs and quasirandomized trials and that the authors reported the risk of bias for each article. More to the point, the article

also presented the risk of bias associated with the body of evidence addressing each treatment effect estimate. For instance, they reported that the direct evidence comparing unreamed with reamed nailing was at low risk of bias, whereas the direct comparison of Ilizarov fixation with external fixation was at high risk of bias owing to inclusion of quasirandomized trials which resulted in a lack of blinding.

Are There Concerns About Publication Bias?

Publication bias occurs when trials that are negative (that is, trials that do not identify a difference between the interventions they evaluated) remain unpublished. Because systematic reviews of the literature are better at identifying published trials than unpublished ones, this can bias estimates of effect sizes upward. Certain patterns of results, referred to as asymmetrical funnel plots, may suggest publication bias [13]. Similarly, a group of small positive studies that are predominantly industry-funded also may suggest publication bias [19]. Neither of these patterns of data is definitive and making inferences regarding publication bias is challenging. Readers should look for evidence that authors have considered the issue by some combination of a search for unpublished data, construction of funnel plots with associated statistical tests, and examination of the pattern of results in relation to industry funding.

Are the Results of Individual Studies in the Direct Comparisons Consistent?

A NMA is based on numerous paired comparisons, each of which typically will involve multiple studies, generating a pooled estimate of effect. In each such comparison, treatment effects will vary, to a lesser or greater extent, across individual studies—we refer to this between-study inconsistency as heterogeneity. Heterogeneity, when substantial, may be attributable to identifiable differences in patients, interventions, outcomes, and/or risk of bias.

Two RCTs evaluating the efficacy of recombinant human BMP-2 (rhBMP-2) help illustrate the importance of this concept. Govender et al. [8] evaluated rhBMP-2 in patients with open tibial shaft fractures and found a large reduction in the need for secondary procedures. Aro et al. [1] conducted a subsequent RCT in a similar population and found no difference in secondary procedures. Aro et al. [1] postulated that the difference was attributable to the type of nailing: the first trial included a majority of patients treated with unreamed nailing whereas the subsequent trial was restricted to patients with reamed nailing. In the trial

by Govender et al. [8], benefit was restricted to patients who had undergone unreamed nailing and the subsequent trial was conducted to clarify the effect in patients undergoing reamed nailing (and it confirmed the lack of benefit seen in in the initial trial).

Clinicians should expect authors of NMAs to note such inconsistencies in direct comparisons and explore possible explanations. Unfortunately, the explanation for heterogeneity in a population of studies can be difficult to discern. When differences in results between studies remain unexplained, our certainty in the pooled estimates that emerge from the analysis decreases.

As with risk of bias, the extent of heterogeneity will differ between paired comparisons—heterogeneity is likely to be small for some comparisons and large for others. To interpret the results of a NMA, one needs to know the degree of heterogeneity associated with each relevant direct comparison.

If a NMA has provided forest plots for each of the direct comparisons, clinicians can assess heterogeneity by examining the extent to which the point estimates are similar and confidence or credible intervals overlap. Statistical tests, such as Cochran's Q test—the statistical test for heterogeneity—or the I^2 statistic can provide additional guidance. The I^2 statistic quantifies the percentage of between-trial variation attributable to between-study differences rather than random error; its value ranges from 0% to 100%. Lower p values in the test for heterogeneity and higher I^2 values are associated with greater heterogeneity and therefore less certainty in effect estimates.

In the open tibial shaft management NMA by Foote et al. [7], only two pairwise comparisons included a sufficient number of studies for a meaningful assessment of heterogeneity. An initial visual assessment of the forest plots comparing unreamed nailing with external fixation (Fig. 1) revealed that odds ratios (ORs) were similar across studies and CIs overlapped; the p value for heterogeneity was 0.83 and $I^2 = 0\%$. The similar point estimates, wide overlap of

CIs, and results of the statistical tests provided no reasons to reduce certainty in the results. In the four trials comparing unreamed with reamed intramedullary nailing, point estimates were similar, CIs had wide overlap, and the p value for the test for heterogeneity was 0.61, with $I^2 = 0\%$, once again showing very consistent results across studies.

Do Indirect Comparisons Respect the Transitivity Principle?

The most straightforward indirect comparisons involve making deductions about the effect of A vs C from the direct comparisons A vs B and B vs C. In Part I of the Users' Guide for Surgeons, we discussed the issue of transitivity—indirect comparisons are secure only if the A vs B and B vs C comparisons are similar with respect to patients and outcomes and the interventions are optimally administered. If there are differences in these factors that influence the magnitude of effect (referred to as effect modifiers), indirect comparisons will be misleading.

Disease severity can act as an effect modifier if, for instance, surgical procedures A and C are, relative to B, superior in less severe injuries but not in more severe injuries. Therefore, if surgical procedure A is tested against B in trials of only Gustilo Type I open tibia fractures, and procedure C is tested against B in trials of only Gustilo Type IIIB open tibia fractures, the indirect comparison will lead to a spurious conclusion that A is superior to C.

Meta-regression is a statistical technique that sometimes can address the presence of effect modifiers and represents an extension of subgroup analysis [14, 16]. Meta-regression attempts to account for effect modification by modeling the treatment effect as a function of the treatment and the effect modifier. In general, potential effect modifiers amenable to meta-regression analysis should be determined a priori to avoid data dredging and spurious inferences of subgroup effects.

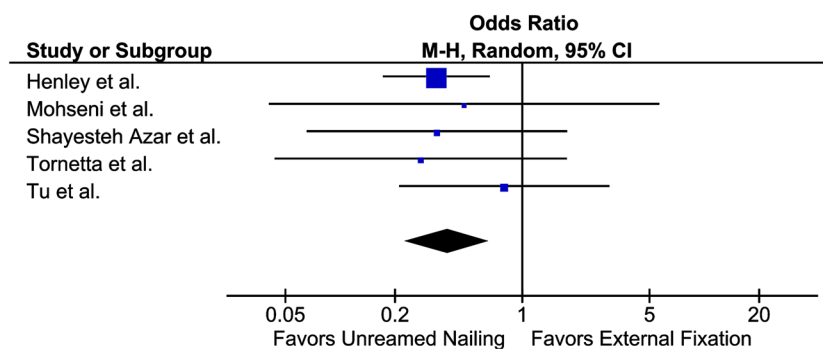


Fig. 1 In the forest plot comparing unreamed intramedullary nailing and external fixation (based on data from the open tibial shaft NMA [7]), initial visual assessment indicates that heterogeneity is low, as odds ratios are similar and CIs are widely overlapping.

Are the Results of Direct and Indirect Comparisons Consistent?

In addition to inconsistent results between individual studies in direct comparisons (heterogeneity), inconsistent results between direct and indirect comparisons may undermine certainty in effect sizes and directions. As introduced in Part I, inconsistency between comparisons is referred to as incoherence. In theory, incoherence usually will result from bias in the indirect comparisons as a result of intransitivity, but incoherence occasionally can be attributable to bias in the relevant direct comparison.

When assessing incoherence, clinicians should look at the magnitude and direction of effect for each paired comparison. The larger the difference between direct and indirect estimates and the less the CIs overlap, the greater the concern regarding incoherence. In addition, NMAs may, and should, present the results of statistical tests to determine the likelihood that chance can explain apparent differences between the two estimates [5, 15]. The lower the associated p value, the greater the concern about incoherence.

When direct and indirect comparisons are coherent, the network estimate will provide a better estimate of the true treatment effect than either the direct or indirect evidence alone. When incoherence exists, which results one should use as the best estimate (direct, indirect, or network) remains controversial [6, 26]. If one body of evidence warrants higher certainty than the other (this will usually but not always be the direct evidence), one can argue strongly that the body of evidence in which there is more certainty represents the best estimate of the true effect [23]. This approach, rather than always using the network estimate as the best estimate, is what the GRADE Working Group suggests.

To illustrate this concept, consider direct, indirect, and combined network outcomes for the pairwise comparison between reamed and unreamed intramedullary nailing as reported by Foote et al. [7]. Direct evidence suggests a lower reoperation rate with unreamed nailing compared with reamed nailing but with a wide CI (OR, 0.74; 95% CI, 0.45–1.24). The indirect evidence, however, suggests a much greater reduction in reoperations with unreamed

nailing compared with reamed nailing (OR, 0.07; 95% CI, 0.01–0.46), with a much narrower CI and almost no overlap with the CI around the direct estimate. Furthermore, a p value is provided for the null hypothesis that the direct and indirect effects are identical; the result (a low p value for the test for incoherence, 0.02) suggests they are not (Table 2).

Because the NMA also used GRADE to assess certainty, we can use the certainty rating to facilitate our interpretation of the results. The indirect evidence was composed from trials at high risk of bias (lacking allocation concealment and blinding) and suspicion of effect modification (intransitivity), and therefore was of low certainty. Because the network estimate was imprecise and based on incoherent data, Foote et al. [7] rated the NMA estimate low certainty as well. However, the direct comparison was rated moderate certainty, and therefore users should trust the direct comparison over the combined network comparison.

Are Confidence or Credible Intervals Sufficiently Narrow?

As noted previously, investigators conducting NMAs will provide, along with best estimates of effect, ranges of results that reflect the degree of random error associated with the estimates. The ranges are known as confidence intervals in conventional frequentist analyses or credible intervals in Bayesian analyses [25].

Treatment networks typically will have an asymmetric configuration, with most of the studies having focused on only a few of the interventions [18]. When multiple trials of sufficient sample size are pooled to directly compare two interventions, CIs are likely to be narrow. If comparative evidence comes largely or completely from indirect estimates, CIs are likely to be wider (although a very large body of indirect evidence can generate narrow CIs). The wider the CI around the NMA estimates of effect, the less certain one can be about the true effect. In a typical NMA, the CIs will be wide for some comparisons and narrow for others, again reinforcing the message that all pairwise comparisons must be individually assessed.

Table 2. Estimates of the impact of unreamed vs reamed intramedullary nailing on reoperation

Comparison	Direct evidence OR (95% CI)	Direct evidence certainty in estimates	Indirect evidence OR (95% CI)	Indirect evidence certainty in estimates	Network OR (95% CI)	Network certainty in estimates
Unreamed compared with reamed	0.74 (0.45 – 1.24)	⊕⊕⊕O MODERATE	0.07 (0.01–0.46)	⊕⊕OO LOW	0.62 (0.37–1.03)	⊕⊕OO LOW

OR = odds ratio; direct, indirect, and network estimates of the effect of unreamed vs reamed intramedullary nailing on reoperation as reported in a network meta-analysis on management strategies for open tibial shaft fractures [7]. In this case, the estimate from direct evidence warrants higher certainty than the combined network estimate.

To decide whether the CI is sufficiently narrow, clinicians must consider what clinical action would be recommended to patients if the upper boundary of the CI represented the truth, and whether that recommendation would change if the lower boundary represented the truth (Table 2).

Focusing first on the estimates from the direct comparison, if, relative to reamed nailing, unreamed nailing reduced the relative odds of a reoperation by 55% (the lower boundary OR of 0.45), one would recommend unreamed nailing [17]. If, however, unreamed nailing increased the relative odds of reoperation by 24% (the upper boundary OR of 1.24), reamed nailing would be the clear choice. Because the choice differs at one end of the CI compared with the other, we would rate down for imprecision.

In contrast, consider the indirect estimate: regardless of whether the true reduction in relative odds of reoperation was 99% (OR of 0.01) or 54% (OR of 0.46), one would recommend unreamed nailing for the patient. Thus, there is no reason to decrease certainty for imprecision (although here certainty is rated down for higher risk of bias and indirectness or intransitivity) [12].

Are the Treatment Rankings Trustworthy?

When a NMA includes a multitude of treatments, deciding on their relative merits by examining the best estimates of effect for each paired comparison (along with confidence in those comparisons) becomes challenging. A simple and attractive solution to the problem is to present this information as a rank order of treatments, from best to worst. Unfortunately, such ranking often can be profoundly misleading.

As each pairwise comparison is associated with certainty and that certainty can range from high to low, it is

important that ranking—or at the very least, the interpretation of ranks—considers the degree of uncertainty. One source of uncertainty is imprecision. Statistical models can take into account uncertainty introduced by imprecision by considering not only ranking, but also the probability of a treatment being ranked 1st, 2nd, 3rd, and so forth. One way to present the information is visually through a “rankogram,” where the horizontal axis represents ascending rank numbers and the vertical axis represents absolute probability [24]. In the instance of a hypothetical NMA of different management options for displaced femoral neck fractures and 1-year functional outcomes (Fig. 2), the appropriate interpretation of the rankogram is not “THA is best” but rather that it remains plausible either THA or bipolar hemiarthroplasty is best; the probability differs little between the two treatments.

Moreover, the rankogram says nothing about risk of bias, inconsistency, indirectness, or the likelihood of publication bias in the trials that contribute to the estimates and thus to the ranking. For example, if outcomes of THA are based on RCTs at high risk of bias while those of bipolar hemiarthroplasty are based on RCTs at low risk of bias, more confidence should be placed in bipolar hemiarthroplasty as the superior procedure even if its statistical rank is lower. Clinicians may need guidance in coming to such appropriate interpretations (and thus, the danger of the ranking exercise).

A second way to present ranking is by reporting the surface under the cumulative ranking curve (SUCRA) [4]. The SUCRA provides a single summary number, where 1 (or 100%) means the treatment is certainly the best and 0 (or 0%) means the treatment is certainly the worst. In the hypothetical example given above, THA may result in a score of 80%, bipolar hemiarthroplasty may result in a score of 60%, and monopolar hemiarthroplasty may result

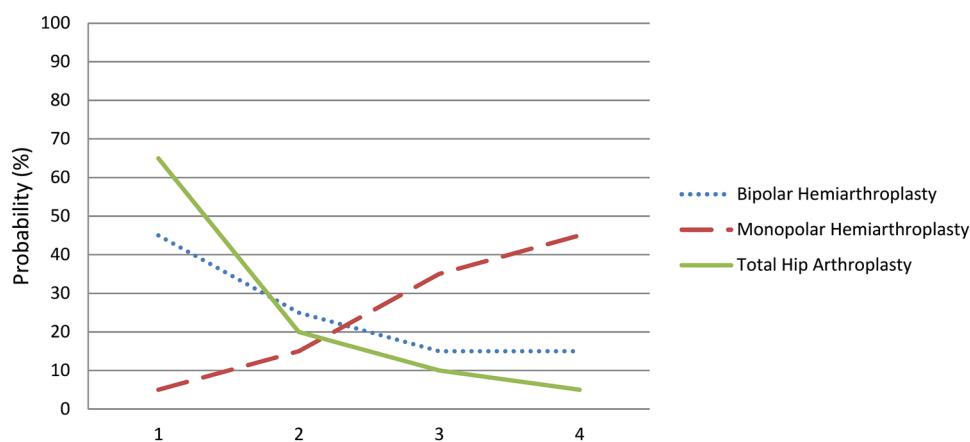


Fig. 2 A rankogram plots the probability of each rank for each treatment. In this hypothetical figure, THA (green solid line) has a higher likelihood of being the best treatment compared with either

monopolar (red dashed line) or bipolar (blue dotted line) hemiarthroplasty. Through visual depiction, ranks are shown as probabilities rather than absolute values.

in a score of 20%. A rank-order list then can be created based on the descending probability scores.

In the open tibial shaft fracture NMA [7], Foote et al. found that direct evidence demonstrated a 60% odds reduction in reoperation with Ilizarov as compared to external fixation. The reported odds reduction from direct evidence was almost as great as that seen between unreamed nailing and external fixation. If Ilizarov fixation was included in the rankings, it would have ranked as the second best treatment. However, the body of evidence for Ilizarov fixation included only one small trial at high risk of bias. Including it in the rankings therefore would be misleading. Foote et al. dealt with this problem by removing treatments from the rankings which had rankings of very low certainty of evidence—this serves as an example of how rankings can be misleading and how authors can avoid misleading rankings.

Box 1. Resolution of Clinical Scenario

Armed with your new critical appraisal skills, you assess the results of the open tibial shaft NMA [7]. Included trials vary in their risk of bias (summarized, using the Cochrane risk of bias tool [modified] [11]). You note that, in general, the evidence base is at high risk of bias: only one of the trials being assessed is assessed as low risk of bias. Further, there are only two treatment comparisons that have multiple direct comparison trials. Heterogeneity is low where there are multiple comparisons; however, most of the direct comparisons are informed by only a single trial. The authors present forest plots for each comparison, along with absolute event rates, ORs, and 95% CIs.

As you assess the results from direct, indirect, and combined network comparisons, you note there is incoherence in some of the comparisons (eg, reamed vs unreamed intramedullary nailing). However, you are able to discern the certainty of treatment effect owing to the authors' use of GRADE assessments. Unfortunately, the highest ranking is moderate, and most are low. The authors present SUCRA score rankings in which unreamed intramedullary nailing is most likely to be the best treatment (SUCRA = 86%), reamed intramedullary nailing is ranked second (SUCRA = 54%), external fixation is third (SUCRA = 24%), and plating is most likely to be the worst treatment (SUCRA = 4%). The authors have excluded trials with very low confidence to avoid misleading rankings, although the low certainty of most estimates undermines the trustworthiness of the rankings.

Based on the information, you conclude that evidence from randomized trials shows that your initial decision to use an unreamed intramedullary nail was likely the best treatment option for preventing reoperation in your patient with an open tibial shaft fracture. You recognize, however, that certainty in all the estimates is moderate or lower. Higher-quality evidence is needed and you are able to see where evidence is lacking.

Conclusion

A NMA is a novel method that enables simultaneous comparison of multiple treatments, taking into account all

existing evidence (direct and indirect). As such, NMAs can be of considerable use to clinicians and their patients, policymakers, and researchers.

As in any other type of study, however, NMAs may be well or poorly done. Even if well done, the evidence may warrant only weak inferences (inferences of low or very low certainty). In assessing this certainty, readers must assess the various risks of bias, ensure the transitivity assumption has not been violated, evaluate consistency between studies and between comparisons, ensure credible intervals or CIs are sufficiently narrow, and interpret rankings with caution (Box 1). This, together with an evaluation of NMA credibility (as presented in Part I of this Users' Guide) will ensure that results of a NMA are interpreted and applied correctly.

References

1. Aro HT, Govender S, Patel AD, Hernigou P, Perera de Gregorio A, Popescu GI, Golden JD, Christensen J, Valentin A. Recombinant human bone morphogenetic protein-2: a randomized trial in open tibial fractures treated with reamed nail fixation. *J Bone Joint Surg Am*. 2011;93:801–808.
2. Bhandari M, Richards RR, Sprague S, Schemitsch EH. Quality in the reporting of randomized trials in surgery: is the Jadad scale reliable? *Control Clin Trials*. 2001;22:687–688.
3. Chess LE, Gagnier J. Risk of bias of randomized controlled trials published in orthopaedic journals. *BMC Med Res Methodol*. 2013;13:76.
4. Cipriani A, Higgins JP, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med*. 2013;159:130–137.
5. Cope S, Donohue JF, Jansen JP, Kraemer M, Capkun-Niggli G, Baldwin M, Buckley F, Ellis A, Jones P. Comparative efficacy of long-acting bronchodilators for COPD: a network meta-analysis. *Respir Res*. 2013;14:100.
6. Druyts E, Thorlund K, Humphreys S, Lion M, Cooper CL, Mills EJ. Interpreting discordant indirect and multiple treatment comparison meta-analyses: an evaluation of direct acting antivirals for chronic hepatitis C infection. *Clin Epidemiol*. 2013;5:173–183.
7. Foote CJ, Guyatt GH, Vignesh N, Mundi R, Chaudhry H, Heels-Ansdell D, Thabane L, Tornetta PIII, Bhandari M. Which surgical treatment for open tibial shaft fracture results in the fewest reoperations? A network meta-analysis. *Clin Orthop Relat Res*. 2015 Feb 28. [Epub Ahead of Print]
8. Govender S, Csimma C, Genant HK, Valentin-Opran A, Amit Y, Arbel R, Aro H, Atar D, Bishay M, Börner MG, Chiron P, Choong P, Cinats J, Courtenay B, Feibel R, Geulette B, Gravel C, Haas N, Raschke M, Hammacher E, van der Velde D, Hardy P, Holt M, Josten C, Ketterl RL, Lindeque B, Lob G, Mathevon H, McCoy G, Marsh D, Miller R, Munting E, Oevre S, Nordsletten L, Patel A, Pohl A, Rennie W, Reynders P, Rommens PM, Rondia J, Rossouw WC, Daneel PJ, Ruff S, Rüter A, Santavirta S, Schildhauer TA, Gekle C, Schnettler R, Segal D, Seiler H, Snowdowne RB, Stapert J, Taglang G, Verdonk R, Vogels L, Weckbach A, Wentzensen A, Wisniewski T; BMP-2 Evaluation in Surgery for Tibial Trauma (BESTT) Study Group. Recombinant human bone morphogenetic protein-2 for treatment of open tibial fractures: a prospective, controlled, randomized study of four hundred and fifty patients. *J Bone Joint Surg Am*. 2002;84:2123–2134.

9. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D, Kunz R, Montori V, Jaeschke R, Rind D, Dahm P, Akl EA, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Schünemann HJ. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol*. 2013;66:151–157.
10. Guyatt GH, Alonso-Coello P, Vandvik PO. Experience with GRADE. *J Clin Epidemiol*. 2012;65:1243–1244.
11. Guyatt GH, Busse JW. Tools to assess risk of bias in randomized controlled trials. Available at: <http://distillercer.com/wp-content/uploads/2014/02/Tool-to-Assess-Risk-of-Bias-in-Randomized-Controlled-Trials.docx>. Accessed February 12, 2015.
12. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW Jr, Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schünemann HJ. GRADE guidelines: 6. Rating the quality of evidence: imprecision. *J Clin Epidemiol*. 2011;64:1283–93.
13. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW Jr, Meerpohl J, Norris SL, Akl EA, Schünemann HJ. GRADE guidelines: 5. Rating the quality of evidence: publication bias. *J Clin Epidemiol*. 2011;64:1277–1282.
14. Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. *BMC Med Res Methodol*. 2012;12:152.
15. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC. 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.
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. Leopold SS. Editor's spotlight/take 5: Comparative responsiveness and minimal clinically important differences for idiopathic ulnar impaction syndrome. *Clin Orthop Relat Res*. 2013;471:1403–1405.
18. Leucht S, Cipriani A, Spineli L, Mavridis D, Orey D, Richter F, Samara M, Barbui C, Engel RR, Geddes JR, Kissling W, Stapf MP, Lässig B, Salanti G, Davis JM. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet*. 2013;382:951–962.
19. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ*. 2003;326:1167–1170.
20. Mills EJ, Bansback N, Ghement I, Thorlund K, Kelly S, Puhan MA, Wright J. Multiple treatment comparison meta-analyses: a step forward into complexity. *Clin Epidemiol*. 2011;3:193–202.
21. Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA*. 2012;308:1246–1253.
22. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6:e1000097.
23. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, Kessels AG, Guyatt GH; GRADE Working Group. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ*. 2014;349:g5630.
24. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64:163–171.
25. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res*. 2001;10:277–303.
26. Thorlund K, Druyts E, Avina-Zubieta JA, Wu P, Mills EJ. Why the findings of published multiple treatment comparison meta-analyses of biologic treatments for rheumatoid arthritis are different: an overview of recurrent methodological shortcomings. *Ann Rheum Dis*. 2013;72:1524–1535.