16

Network Meta-analysis

Edward J. Mills, John P. A. Ioannidis, Kristian Thorlund, Holger J. Schünemann, Milo A. Puhan, and Gordon Guyatt

IN THIS CHAPTER

(continued on following page)

CLINICAL SCENARIO

Your patient is a 45-year-old woman who experiences frequent migraine headaches that last from 4 to 24 hours and prevent her from attending work or looking after her children. She has exhausted efforts to manage the symptoms with nonsteroidal anti-inflammatory drugs and seeks additional treatment. You decide to recommend a triptan for the patient's migraine headaches but are wondering how to choose from the 7 available triptans. You retrieve a *network meta-analysis* (NMA) that evaluates the different triptans among this patient population.[1] You are not familiar with this type of study, and you wonder if there are special issues to which you should attend in evaluating its methods and results.

FINDING THE EVIDENCE

You start by typing "migraine triptans" in the search box of an evidence-based summary website with which you are familiar. You find several chapters related to the management of migraine and drug information on the different drugs that are available. However, despite the profusion of *evidence* comparing single regimens, you wonder if all triptans have been compared, ideally in a in a single *systematic review*. To search for such a review, you type "migraine triptans comparison" in PubMed's Clinical Queries (http://www.ncbi.nlm.nih.gov/pubmed/clinical; see Chapter 4, Finding Current Best Evidence). In the results page, the middle column, which applies a broad filter for potential systematic reviews, retrieves 21 citations. The first strikes you as the most relevant to your question. It is a network meta-analysis that evaluates the different triptans among your patient population.[1] You are not familiar with this type of study, and you wonder if there are special issues to which you should attend in evaluating its methods and results.

INTRODUCTION

Traditionally, a meta-analysis addresses the merits of one intervention vs another (eg, *placebo* or another active intervention). Data are combined from all studies—often *randomized clinical trials* (RCTs)—that meet eligibility criteria in what we will term a pairwise meta-analysis. Compared with a single RCT, a *meta-analysis* improves the power to detect differences and also facilitates examination of the extent to which there are important differences in *treatment effects* across eligible RCTs—variability that is frequently called *heterogeneity*.[2,3] Large unexplained heterogeneity may reduce a reader's confidence in estimates of treatment effect (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

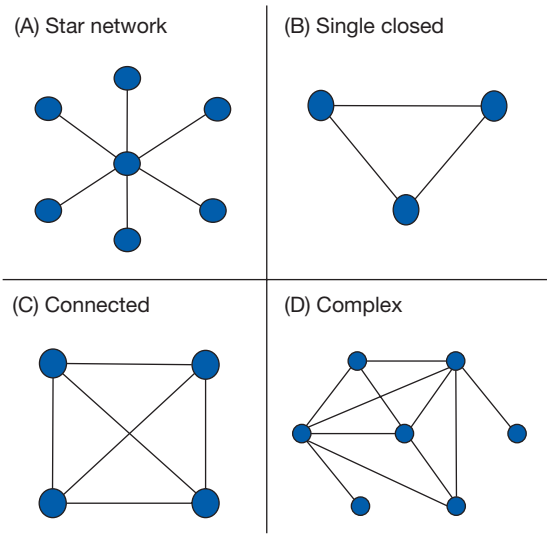
A drawback of traditional pairwise meta-analysis is that it evaluates the effects of only 1 intervention vs 1 comparator and does not permit inferences about the relative effectiveness of several interventions. For many medical conditions, however, there are a selection of interventions that have most frequently been compared with placebo and occasionally with one another.[4,5] For example, despite 91 completed and ongoing RCTs that address the effectiveness of the 9 biologic drugs for the treatment of rheumatoid arthritis, only 5 compare biologics directly against each other.[4]

Recently, another form of meta-analysis, called an NMA (also known as a multiple or mixed treatment comparison meta-analysis) has emerged.[6,7] The NMA approach provides estimates of effect sizes for all possible pairwise comparisons whether or not they have actually been compared head to head in RCTs. Figure 16-1 displays examples of common networks of treatments.

Our ability to provide estimates of relative effect when 2 interventions, A and B, have not been tested head to head against one another comes from what are called indirect comparisons. We can make an indirect comparison if the 2 interventions

FIGURE 16-1

Examples of Possible Network Geometry



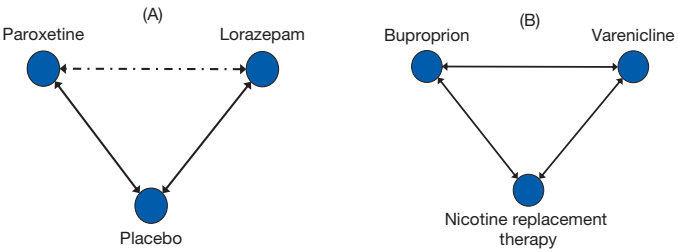
The figure shows 4 network graphs. In each graph, the lines show where direct comparisons exist from 1 or more trials. Figure 16-1A shows a star network, where all interventions have just 1 mutual comparator. Figure 16-1B shows a single closed loop that involves 3 interventions and can provide data to calculate both direct comparisons and indirect comparisons. Figure 16-1C shows a well-connected network, where all interventions have been compared against each other in multiple randomized clinical trials. Figure 16-1D is an example of a complex network with multiple loops and also arms that have sparse connections.

(eg, paroxetine and lorazepam in Figure 16-2A) have each been compared directly against another intervention, C (eg, placebo).

For instance, assume that A (eg, paroxetine) substantially reduces the odds of an adverse outcome relative to C (placebo) (*odds ratio* [OR], 0.5). Intervention B (eg, lorazepam), on the other hand, has no impact relative to C on that outcome (OR, 1.0).

FIGURE 16-2

A Simple Indirect Comparison and Simple Closed Loop



In the first example (A), there is direct evidence from paroxetine compared with placebo and direct evidence of lorazepam compared with placebo. Therefore, the indirect comparison can be applied to determine the effect of paroxetine compared with lorazepam, even if no direct head-to-head comparison exists on these 2 agents. In the second example (B), there is direct evidence that compares nicotine replacement therapy with both varenicline and bupropion. There is also direct evidence that compares bupropion with varenicline. Therefore, one has enough information to evaluate whether the results are coherent between direct and indirect evidence.

One might then reasonably deduce that A is substantially superior to C—indeed, our best estimate of the OR of A vs B would be 0.5/1.0 or 0.5. The ratio of the OR in such a situation is our way of estimating the effect of A vs B on the outcome of interest.[8]

Network meta-analyses, which simultaneously include both direct and *indirect evidence* (see Figure 16-2B for an example in which both direct and indirect evidence is available, sometimes called a closed loop), are subject to 3 chief considerations. The first is an assumption that is also necessary for a conventional meta-analysis (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). Among trials available for pairwise comparisons, are the studies sufficiently *homogenous* to combine for each intervention?

Second, are the trials in the network sufficiently similar, with the exception of the intervention (eg, in important features, such as populations, design, or outcomes)?[9] For instance, if trials of drug

A vs placebo differ substantially in the characteristics of the population studied from the population in drug B vs placebo, inferences about the relative effect of A and B on the basis of how each did against placebo become questionable. Third, where direct and indirect evidence exist, are the findings sufficiently consistent to allow confident pooling of direct and indirect evidence together?

By including evidence from both direct and indirect comparisons, an NMA may increase precision of estimates of the relative effects of treatments and facilitate simultaneous comparisons, or even ranking, of these treatments.[7] However, because NMAs are methodologically sophisticated, they are often challenging to interpret.[10]

One challenge clinicians will face with NMAs is that they usually use *Bayesian analysis* approaches rather than the *frequentist analysis* approaches with which most of us are more familiar. Clinicians need not worry further about this, and the main reason for pointing it out is as an alert to a difference in terms. Clinicians are used to considering *confidence intervals* (CIs) around estimates of treatment effect. The Bayesian equivalent are called *credible intervals* and can be interpreted in conceptually the same way as CIs.

Here, we demystify NMAs by using the 3 questions of *risk of bias*, results, and applicability of results. Box 16-1 includes all issues relevant to evaluating systematic reviews. Our discussion in this chapter does not include all of the issues but rather highlights those that are most important, or differ, in NMAs.

BOX 16-1

Users' Guides Critical Appraisal Tool

How serious is the risk of bias?

Did the review include explicit and appropriate eligibility criteria?

Was biased selection and reporting of studies unlikely?

Did the review address possible explanations of between-study differences in results?

Were selection and assessments of studies reproducible?

Did the authors rate the confidence in effect estimates for each paired comparison?

What are the results?

What was the amount of evidence in the network?

Were the results similar from study to study?

Were the results consistent in direct and indirect comparisons?

How did treatments rank and how secure are we in the rankings?

Were the results robust to sensitivity assumptions and potential biases?

How can I apply the results to patient care?

Were all patient-important outcomes considered?

Were all potential treatment options considered?

Are any postulated subgroup effects credible?

What is the overall quality and what are limitations of the evidence?

HOW SERIOUS IS THE RISK OF BIAS?

Did the Meta-analysis Include Explicit and Appropriate Eligibility Criteria?

One can formulate questions of optimal patient management in terms of the *PICO* framework of patients (P), interventions (I), comparisons (C), and outcomes (O).

Broader eligibility criteria may enhance *generalizability* of the results but may be misleading if participants are too dissimilar and as a consequence heterogeneity is large. Diversity of interventions may also be excessive if authors pool results from

different doses or even different agents in the same class (eg, all statins), based on the assumption that effects are similar. You should ask whether investigators have been too broad in their inclusion of different populations, different doses or different agents in the same class, or different outcomes to make comparisons across studies credible.

Was Biased Selection and Reporting of Studies Unlikely?

Some NMAs apply the search strategies from other systematic reviews as the basis for identifying potentially eligible trials. Readers can be confident in such approaches only if authors have updated the search to include recently published trials.[11]

The eligible interventions can be unrestricted. Sometimes, however, the authors may choose to include only a specific set of interventions, eg, those available in their country. Some industry-initiated NMAs may choose to consider only a sponsored agent and its direct competitors.[12] This may omit the optimal agent for some situations and tends to give a fragmented picture of the evidence. It is typically best to include all interventions[13] because data on clearly suboptimal or abandoned interventions may still offer indirect evidence for other comparisons.[14]

In an NMA of 12 treatments for major depression, the authors chose to exclude placebo-controlled RCTs and included only head-to-head active treatment RCTs.[15] However, *publication bias* in the antidepressant literature is well acknowledged,[16,17] and by excluding placebo-controlled trials, the analysis loses the opportunity to benefit from additional available evidence.[18] Exclusion of eligible interventions, in this case placebo, may not just decrease statistical power but may also change the overall results.[14] Placebo-controlled trials may be different

than head-to-head comparison trials in their conduct or in the degree of bias (eg, they may have more or less publication bias or *selective outcome reporting* and selective analysis reporting). Thus, their exclusion may also have an impact on the *point estimates* of the effects of pairwise comparisons and may affect the relative ranking of regimens.[14] When an NMA of second-generation antidepressants was later conducted and included placebo-controlled trials, relying only on the relative differences among treatments using the same depression scale, the authors reached a different interpretation than the earlier NMA.[15,19,20]

Finally, original trials often address multiple outcomes. Selection of NMA outcomes should not be data driven but based on importance for patients and consider both outcomes of benefit and *harm*.

Did the Meta-analysis Address Possible Explanations of Between-Study Differences in Results?

When substantial clinical variability is present (this is usually, and appropriately, the case), authors may conduct *subgroup analyses* or *meta-regression* to explain heterogeneity. If such analyses are successful in explaining heterogeneity, the NMA may provide results that more optimally fit the clinical setting and characteristics of the patient you are treating.[21] For example, in an NMA evaluating different statins for cardiovascular disease protection, the authors used meta-regression to address whether it was appropriate to combine results across primary and secondary prevention populations, different statins, and different doses of statins.[22] Meta-regression suggested heightened efficacy in those with prior cardiac events and those with a

history of hypertension, possibly suggesting a more compelling case for statin use in such populations.

Inclusion of multiple control interventions (eg, placebo, no intervention, older standard of care) may enhance the robustness and connectedness of the treatment network. It is, however, important to gauge and account for potential differences between control groups. For example, because of potential placebo effects, patients receiving placebo in a *blinded* RCT may have differing responses than patients receiving no intervention in a nonblinded RCT. Thus, if an active treatment, A, has been compared with placebo and another active treatment, B, has been compared with no intervention, the different choice of control groups may produce misleading results (B may appear superior, but the use of placebo as the comparator in the A trials may be responsible for the difference). As with active interventions, meta-regression may address this problem.

For example, in an NMA evaluating the effectiveness of smoking cessation therapies, the authors combined placebo-controlled arms with standard-of-care control arms and then used meta-regression to examine whether the choice of control changed the *effect size*.[23] The authors found that trials that used placebo controls had smaller effect sizes than those that used standard of care, which explained the heterogeneity.

Did the Authors Rate the Confidence in Effect Estimates for Each Paired Comparison?

The treatment effects in an NMA are typically reported with common effect sizes along with 95% credible intervals. Credible intervals are the Bayesian equivalent to the more commonly understood CIs. When there are K interventions included in the treatment network, there are $K^*(K-1)/2$ possible pairwise

comparisons. For example, if there are 7 interventions, then there are 7*(7–1)/2, or 21, possible pairwise comparisons. Like authors of conventional meta-analyses, authors of NMAs need to address confidence in estimates of effect for each paired comparison (A vs B, A vs C, B vs C, etc—15 comparisons in the NMA example with 7 interventions). The necessity for these confidence ratings is that evidence may warrant strong inferences (ie, high confidence in estimates) for the superiority of one treatment over another (A vs B, for instance) and only weak inferences (ie, very low confidence in estimates) for the judgment of superiority of another pairing (A vs C).

The GRADE Working Group has provided a framework that is well suited to addressing confidence in estimates (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). We lose confidence in direct comparisons of alternative treatments if the relevant randomized trials have failed to protect against risk of bias by *allocation concealment*, blinding, and preventing loss to *follow-up* (see Chapter 6, Therapy [Randomized Trials]). We also lose confidence when CIs (or in the case of a Bayesian NMA, credible intervals) on pooled estimates are wide (imprecision); results vary from study to study and we cannot explain the differences (*inconsistency*); the population, intervention, or outcome differ from that of primary interest (*indirectness*); or we are concerned about publication bias.

Ideally, for each paired comparison, authors will present the pooled estimate for the direct comparison (if there is one) and its associated rating of confidence, the indirect comparison(s) that contributed to the pooled estimate from the NMA and its associated rating of confidence, and the NMA estimate and the associated rating of confidence. Criteria for judging confidence in estimates for direct comparisons are well established. Although these criteria provide considerable guidance in assessing confidence in indirect estimates, judgments regarding confidence in estimates from indirect comparisons present additional challenges. Criteria for addressing these challenges are still evolving, reflecting that NMA is still a very new method.

USING THE GUIDE

Returning to our opening scenario, the NMA we identified compared the efficacy of different triptans for the abortive treatment of migraine headaches.[1] Patients of interest included adults 18 to 65 years old who experience migraines, with or without aura. Experimental and control interventions included available oral triptans, placebos, and no-treatment controls. The outcomes of interest were pain-free response at 2 hours and 24 hours after the onset of headache. Patients in the included RCTs met similarly broad diagnostic criteria based on criteria from the International Headache Society and had to experience at least 1 migraine headache every 6 weeks. The outcomes assessed are important to patients, and their definitions were consistent across trials. Moreover, the authors planned to assess dose as a potential effect modifier.

The authors conducted a comprehensive search for published literature and sought unpublished RCTs via contact with industry trialists. Two reviewers conducted the search and extracted data independently, in duplicate. The authors did not rate the confidence in estimates from paired comparisons but provided information that allows conclusions about confidence. The authors reported events as proportions with ORs for treatment effects.

WHAT ARE THE RESULTS?

What Was the Amount of Evidence in the Treatment Network?

One can gauge the amount of evidence in the treatment network from the number of trials, total sample size, and number of events for each treatment and comparison. Furthermore, the

extent to which the treatments are connected in the network is an important determinant of the confidence we can have in the estimates that emerge from the NMA. Understanding the *geometry of the network* (nodes and links) will permit clinicians to examine the larger picture and see what is compared to what.[24] Authors will generally present the structure of the network (as in the examples in Figure 16-1).

When alternative interventions have been compared with a single common comparator (eg, placebo), we call this a star network (Figure 16-1A). A star network only allows for indirect comparisons among active treatments, which reduces confidence in effects, particularly if there are a limited number of trials, patients, and events.[25] When there are data available that use both direct and indirect evidence of the same interventions, we refer to this as a closed loop (Figure 16-1B). The presence of direct evidence increases our confidence in the estimates of interest.

Often, a treatment network will include a mixture of exclusively indirect links and closed loops (Figure 16-1C and D). Most networks have unbalanced shapes with many trials of some comparisons, but few or none of others.[24] In this situation (and indeed, in many situations, as we have pointed out in our discussion of the need for a confidence rating of each paired comparison), evidence may warrant high confidence for some treatments and comparisons but low confidence for others. The credible intervals around direct, indirect, and NMA estimates provide a helpful index of the amount of information available for each paired comparison.

Were the Results Similar From Study to Study?

In a traditional meta-analysis of paired treatment comparisons, results often vary from study to study. Investigators can address possible explanations of differences in treatment effects using a subgroup analysis and meta-regression. However, these analyses are limited in the presence of small numbers of trials, and apparent subgroup effects often prove spurious, an issue to which we return in our discussion of applicability.[26-28]

Network meta-analyses, with larger numbers of patients and studies, present opportunities for more powerful exploration of explanations of between-study differences. Indeed, as we have pointed out in a prior section of this chapter—Did the Review Address Possible Explanations of Between-Study Differences in Results?—the search conducted by NMA authors for explanations for heterogeneity may be informative.

Nevertheless, as is true for conventional meta-analyses, NMA is vulnerable to unexplained differences in results from study to study. Ideally, NMA authors will, in summarizing the results of each paired comparison, alert you to the extent of inconsistency in results in both the direct and indirect comparisons and the extent to which confidence in estimates decreases accordingly (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

Were the Results Consistent in Direct and Indirect Comparisons?

Direct comparisons of treatments are generally more trustworthy than indirect comparisons. However, these head-to-head trials can also yield misleading estimates (eg, when conflicts of interest influence the choice of comparators used or result in selective reporting). Therefore, indirect comparisons may on occasion provide more trustworthy estimates.[29]

Deciding what estimates are most trustworthy (direct, indirect, or network) requires assessing whether the direct and indirect estimates are consistent or discrepant. One can assess whether direct and indirect estimates yield similar effects whenever there is a closed loop in the network (as in Figure 16-2B). Statistical methods exist for checking this type of inconsistency, typically called a test for *incoherence*.[30,31]

A group of investigators applied a test of incoherence to 112 interventions in which direct and indirect evidence was available. They found that the results were statistically inconsistent 14% of the time.[9] This same evaluation found that comparisons

with smaller number of trials and measuring subjective outcomes had a greater risk of incoherence.

Authors' presentation of direct and indirect estimates for each paired comparison will allow you to easily examine the extent of incoherence between direct and indirect estimates. Authors can perform statistical tests to determine whether chance can explain the difference between direct and indirect estimates. Often, however, the amount of data is limited and not sufficient, and important differences may still exist in the absence of a statistically significant difference.

When incoherence is present, there are many explanations for the authors—and for you—to consider (Box 16-2). Just as unexplained heterogeneity in any direct paired comparison decreases confidence in the pooled estimate, unexplained incoherence reduces confidence in the estimate that arises from the network. Indeed, when large incoherence is present, the more credible estimate may come from either the direct (usually) or indirect (seldom) comparison rather than from the network.

BOX 16-2

Potential Reasons for Incoherence Between the Results of Direct and Indirect Comparisons

Chance

Genuine differences in results

- Differences in enrolled participants (eg, entry criteria, clinical setting, disease spectrum, baseline risk, selection based on prior response)
- Differences in the interventions (eg, dose, duration of administration, prior administration [second-line treatment])
- Differences in background treatment and management (eg, evolving treatment and management in more recent years)
- Differences in definition or measurement of outcomes

Bias in head-to-head (direct) comparisons

Optimism bias with unconcealed analysis

Publication bias

Selective reporting of outcomes and of analyses

Inflated effect size in *stopped early trials* and in early evidence

Limitations in allocation concealment, blinding, loss to follow-up, analysis as randomized

Bias in indirect comparisons

Each of the biasing issues above can affect the results of the direct comparisons on which the indirect comparisons are based

For example, a meta-analysis examining the analgesic efficacy of paracetamol plus codeine in surgical pain found a direct comparison that indicated the intervention was more efficacious than paracetamol alone (mean difference in pain intensity change, 6.97; 95% CI, 3.56-10.37). The *adjusted indirect comparison* did not find a significant difference between paracetamol plus codeine and paracetamol alone (–1.16; 95% CI, –6.95 to 4.64).[32] In this example, the direct and indirect evidence was statistically significantly incoherent (P = .02). The explanation for incoherence may be that the direct trials included patients with lower pain intensity at baseline, and such patients may be more responsive to the addition of codeine.

How Did Treatments Rank and How Confident Are We in the Ranking?

Besides presenting treatment effects, authors may also present the probability that each treatment is superior to all other treatments, allowing ranking of treatments.[33,34] Although this

approach is appealing, it may be misleading because of fragility in the rankings, because differences among the ranks may be too small to be important, or because of other limitations in the studies (eg, risk of bias, inconsistency, indirectness).

We have already provided one example of such a misleading ranking: in an NMA of drug treatments to prevent fragility hip fractures, the authors' conclusion that teriparatide had the highest probability of being ranked first across 10 treatments[24] was misleading because comparison of teriparatide with all other agents, including placebo, warranted only low or very low confidence.

In another example, an NMA that examined direct-acting agents for hepatitis C found no statistical difference for sustained virologic response between telaprevir and boceprevir (OR, 1.42; 95% credible interval, 0.89-2.25); on the basis of these results, the probability of being the best favors teleprevir by far (93%) over boceprevir (7%).[35,36] However, this 93% probability provides a misleadingly strong endorsement for teleprevir. The lower boundary of the credible interval tells us that our confidence in substantial superiority of teleprevir is very low.

Examination of the confidence in estimates from each paired comparisons provides insight into the trustworthiness of any rankings, and reveals the importance of providing such ratings.

Were the Results Robust to Sensitivity Assumptions and Potential Biases?

Given the complexity of some NMAs, authors may assess the robustness of their study findings by applying sensitivity

analyses that reveal how the results change if some criteria or assumptions change. Sensitivity analyses may include restricting the analyses to trials with low risk of bias only or examining different but related outcomes. The Cochrane Handbook provides a discussion of sensitivity analyses.[37]

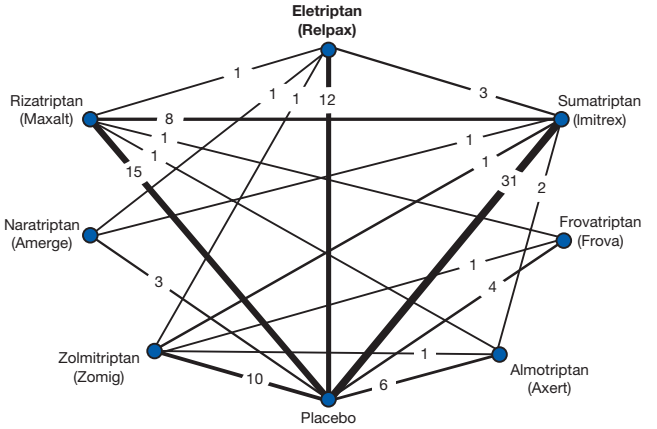
For example, in an NMA on prevention of chronic obstructive pulmonary disease (COPD) exacerbations, the authors used the incidence rate as the primary outcome. However, there is some debate on whether incidence rates should be used in COPD trials,[38] and so the authors conducted sensitivity analyses with the binary outcome of ever having an exacerbation. The results were sufficiently similar to consider the analyses robust.[39]

USING THE GUIDE

Returning to our clinical scenario, Figure 16-3 displays the network of considered treatments for pain-free response at 2 hours. The authors included 74 RCTs that examined triptans for the treatment and prevention of migraine attacks. Placebo was compared with eletriptan, sumatriptan, rizatriptan, zolmitriptan, almotriptan, naratriptan, and frovatriptan in 15, 30, 16, 5, 9, 5, and 4 trials, respectively. The amount of evidence varied across these comparisons. For example, naratriptan had only been compared with placebo in 2 trials; therefore, confidence in these estimates is likely to be low. Evidence for sumatriptan and rizatriptan was based on a larger amount of evidence from both direct and indirect comparisons. Sumatriptan (n = 30), rizatriptan (n = 20), and eletriptan (n = 16) had the most links, whereas placebo was the most connected node (n = 68). The most common

FIGURE 16-3

Treatment Network for the Drugs Considered in the Example Network Meta-analysis on Triptans for the Abortive Treatment of Migraine for Pain-Free Response at 2 Hours



The lines between treatment nodes indicate the comparisons made throughout randomized clinical trials (RCTs). The numbers on the lines indicate the number of RCTs informing a particular comparison.

direct comparisons (n = 4 trials) were between sumatriptan and rizatriptan (the 2 most commonly tested treatments). Of these, 15 comparisons were informed direct evidence, but 7 of the direct connections had only 1 trial, and several of the comparisons were informed only by indirect evidence. Frovatriptan was poorly connected to other treatments, and all comparisons that involved this agent warranted, therefore, only moderate confidence at best.

Sixty-three trials reported the outcome of pain-free response at 2 hours, and 25 reported 24 hours of sustained pain-free response. The authors used the I^2 value to assess heterogeneity in pairwise meta-analysis before conducting their NMA; however, they did not report the specific values. They checked the coherence between direct and indirect comparisons from closed loops and provided this information as a supplemental appendix online. Direct and indirect evidence were consistently similar, with no statistical evidence of incoherence (Table 16-1). The authors also conducted several sensitivity analyses to assess the role of dose.

Figure 16-4 displays the results of the NMA of triptans vs placebo. For pain-free response at 2 hours, the authors found that eletriptan, sumatriptan, and rizatriptan exhibited the largest treatment effects against placebo. The results were largely similar for pain-free response at 24 hours.

When the authors examined the comparative effectiveness of each triptan vs the other triptans, evidence warranted at least moderate confidence for some differences among triptans. For example, eletriptan was superior in pain-free response at 2 hours compared with sumatriptan (OR, 1.53; 95% CI, 1.16-2.01), almotriptan (OR, 2.03; 95% CI, 1.38-2.96), zolmitriptan (OR, 1.46; 95% CI, 1.02-2.09), and naratriptan (OR, 2.95; 95% CI, 1.78-4.90).

For all but naratriptan, we have at least moderate confidence in treatment effects vs placebo at 2 and 24 hours. Eletriptan was associated with the largest probability (68%) of being the best treatment for pain-free response at 2 and 24 hours (54.1%). The only other drug that ranked favorably was rizatriptan (22.6% at 2 hours and 9.2% at 24 hours). Given that comparisons between eletriptan and a number of other agents warrant at least moderate confidence, the first rank of elitriptan carried considerable weight.

TABLE 16-1

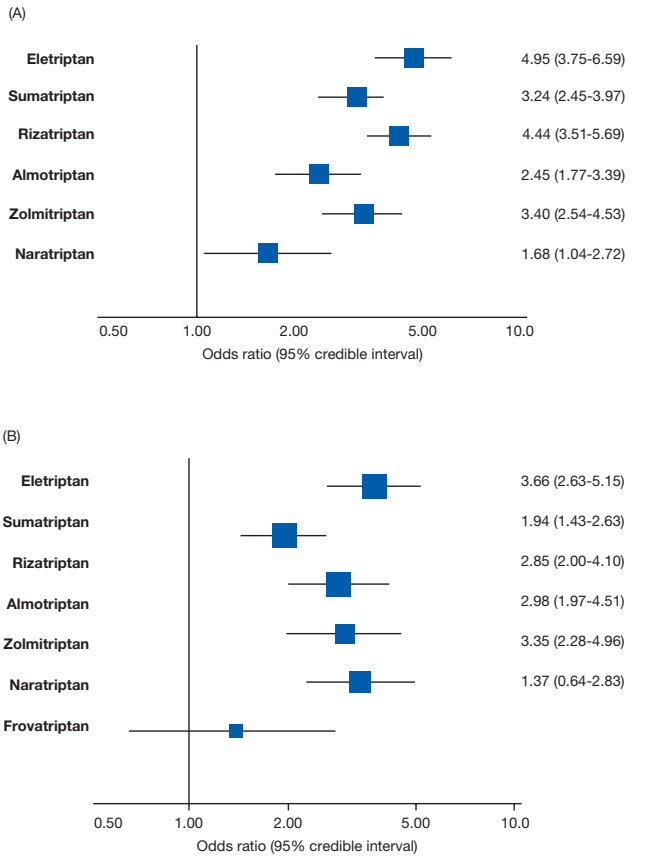
Consistency Check for a Pain-Free Response at 2 Hours With Triptan in Usual Doses

Comparison	No. of Trials	Direct Estimate[a]	Indirect Estimate[a]
Three-Treatment Loops Where Inconsistency Can be Checked			
Eletriptan (40 mg) vs sumatriptan (50 mg)	2	1.48 (1.14-2.79)	1.58 (0.60-5.87)
Eletriptan (40 mg) vs zolmitriptan (12.5 mg)	2	1.52 (0.96-1.81)	1.21 (0.35-3.55)
Eletriptan (40 mg) vs naratriptan (2.5 mg)	1	2.46 (1.53-3.98)	2.75 (0.37-19.8)
Sumatriptan (50 mg) vs almotriptan (2.5 mg)	1	1.49 (1.12-1.98)	1.07 (0.63-1.76)
Sumatriptan (50 mg) vs zolmitriptan (12.5 mg)	1	1.12 (0.87-1.45)	0.72 (0.42-1.29)
Sumatriptan (50 mg) vs frovatriptan (2.5 mg)	1	1.07 (0.56-2.04)	0.64 (0.35-1.15)
Almotriptan (2.5 mg) vs zolmitriptan (12.5 mg)	1	0.89 (0.69-1.15)	0.70 (0.41-1.19)
Zolmitriptan (12.5 mg) vs frovatriptan (2.5 mg)	1	0.73 (0.52-1.02)	0.86 (0.47-1.62)
Naratriptan (12.5 mg) vs frovatriptan (2.5 mg)	1	0.82 (0.51-1.20)	0.90 (0.49-1.79)

[a]Odds ratio estimates and 95% confidence intervals for all treatment comparisons from the direct pairwise meta-analysis of head-to-head trials and indirect comparison meta-analysis using placebo as the common comparator.

FIGURE 16-4

Forest Plot of the Primary Multiple-Treatment Comparison Meta-analysis Results, Triptans vs Placebo



A, Pain-free response at 2 hours; B, 24 hours of sustained pain-free response.

HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

Were All Patient-Important Outcomes Considered?

Many NMAs report only 1 or a few outcomes of interest. For example, a recent NMA that compared the efficacy of antihypertensive treatments reported only heart failure and mortality,[40] whereas an older NMA of antihypertensive treatments also considered coronary heart disease and stroke.[41] Adverse events are infrequently assessed in meta-analysis and in NMAs, reflecting poor reporting in the *primary studies*.[42,43] Network meta-analyses conducted in the context of health technology assessment submissions and *evidence-based practice* reports are more likely to include multiple outcomes and assessments of harms than the less lengthy NMAs published in clinical medical journals.[20]

USING THE GUIDE

The authors assessed outcomes (pain-free response at 2 and 24 hours) that are important to patients. The major omission is adverse events—if triptans differed substantially in adverse events, this would be an important consideration for patients. Fortunately, the drug that appears as or more effective than other triptans, eletriptan, also appears to be at least as well tolerated as other triptans.[44]

Were All Potential Treatment Options Considered?

Network meta-analyses may place restrictions on what treatments are examined. For example, for irritable bowel syndrome, an NMA may focus on pharmacologic agents, neglecting RCTs of diet, peppermint oil, and counseling.[45] Decisions to focus on subclasses of drugs may also be problematic. For example, in

rheumatoid arthritis, biologics are used for patients in whom conventional drugs fail. Five of the 9 available biologics are anti–tumor necrosis factor (TNF) agents. One recent NMA only considered anti-TNF agents and excluded other biologics.[46] To the extent that the other biologic agents are equivalent or superior to the anti-TNF agents, their exclusion risks misleading clinicians regarding the best biologic agents.

Are Any Postulated Subgroup Effects Credible?

There are very few situations in which investigations have convincingly established important differences in the relative effect of treatment according to patient characteristics.[47] Criteria exist for determining the credibility of subgroup analyses.[47] These criteria include whether the comparisons are within-study (subgroup A and subgroup B both participated in the same study, the stronger comparison) or between-study (one study enrolled subgroup A and another subgroup B, the weaker comparison), chance is an unlikely explanation of the differences in effect between subgroups, and the investigators made a small number of a priori subgroup hypotheses with an accurately specified direction. Network meta-analyses allow a greater number of RCTs to be evaluated and may offer more opportunities for subgroup analysis—but with due skepticism and respect for credibility criteria.

For example, in an NMA that examined inhaled drugs for COPD, the authors examined whether severity of airflow obstruction measured by forced expiratory volume in 1 second (FEV_1) influenced patients' response.[48] If the FEV_1 was 40% or less of predicted, long-acting anticholinergics, inhaled corticosteroids, and combination treatment, including inhaled corticosteroids, reduced exacerbations significantly compared with long-acting β-agonists alone

but not if the FEV_1 was greater than 40% of predicted. This difference was significant for inhaled corticosteroids ($P = .02$ for interaction) and combination treatment ($P = .01$) but not for long-acting anticholinergics ($P = .46$). The fact that these analyses were based on an a priori hypothesis, including a correctly hypothesized direction with a strong biologic rationale (greater inflammation in more severe airway disease) and a low *P* value for the test of interaction (ie, chance is an unlikely explanation), strengthens the credibility of the subgroup effect. It is, however, based on a between-group comparison. A reasonable judgment would be moderate to high credibility of the subgroup effect, and a clinical policy of restricting inhaled corticosteroid use to patients with more severe airflow obstruction.

CLINICAL SCENARIO RESOLUTION

You conclude that there is convincing evidence for the role of triptans in aborting migraine headaches at 2 and 24 hours. However, because triptans are a class of drugs you choose to assess whether this *class effect* is real or not. There are data available from direct and indirect comparisons that suggest that eletriptan is superior to several other triptans. You opt to discuss with the patient the benefits of starting treatment with eletriptan and will seek evidence for adverse events.

CONCLUSION

Although an NMA can provide extremely valuable information in choosing among multiple treatments offered for the same condition, it is important to determine the confidence one can

place in the estimates of effect of the treatments considered and the extent to which that confidence differs across comparisons. If authors provide these confidence ratings themselves using criteria such as those suggested by *GRADE* (*Grading of Recommendations Assessment, Development and Evaluation*), the task is straightforward—simply survey the confidence ratings. Those rated as high or moderate are trustworthy and those rated low or very low much less so. If the authors do not provide these ratings themselves, you need to make your own assessments, which can be challenging.

The confidence for any comparison will be greater if individual studies are at low risk of bias and publication bias is unlikely; results are consistent in individual direct comparisons and individual comparisons with no-treatment controls and also consistent between direct and indirect comparisons; sample size is large and CIs are correspondingly narrow; and most comparisons have some direct evidence. If all of these hallmarks are present and the differences in effect sizes are large, high confidence in estimates may be warranted. However, in most cases, confidence in some key estimates is likely to warrant only moderate or low confidence. Most concerning, if authors do not provide the necessary information, it is difficult to judge which comparisons are trustworthy and which less so—and in such cases, clinicians may be best served by reviewing systematic reviews and meta-analyses of the direct comparisons and using these to guide their patient management.

References

1. Thorlund K, Mills EJ, Wu P, et al. Comparative efficacy of triptans for the abortive treatment of migraine: a multiple treatment comparison meta-analysis. *Cephalalgia*. 2014;34(4):258-267.
2. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351(9096):123-127.
3. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med*. 1987;316(8):450-455.

4. Estellat C, Ravaud P. Lack of head-to-head trials and fair control arms: randomized controlled trials of biologic treatment for rheumatoid arthritis. *Arch Intern Med.* 2012;172(3):237-244.
5. Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. *Eur J Clin Invest*. 2010;40(2):172-182.
6. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17(3):279-301.
7. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004;23(20):3105-3124.
8. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
9. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ*. 2011;343:d4909.
10. Mills EJ, Bansback N, Ghement I, et al. Multiple treatment comparison meta-analyses: a step forward into complexity. *Clin Epidemiol*. 2011;3:193-202.
11. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*. 2009;151(4):W65-W94.
12. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics*. 2008; 26(9):753-767.
13. Kyrgiou M, Salanti G, Pavlidis N, Paraskevaidis E, Ioannidis JP. Survival benefits with diverse chemotherapy regimens for ovarian cancer: meta-analysis of multiple treatments. *J Natl Cancer Inst.* 2006;98(22):1655-1663.
14. Mills EJ, Kanters S, Thorlund K, Chaimani A, Veroniki AA, Ioannidis JP. The effects of excluding treatments from network meta-analyses: survey. *BMJ.* 2013;347:f5195.
15. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*. 2009;373(9665):746-758.
16. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008;358(3):252-260.
17. Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? *Philos Ethics Humanit Med*. 2008;3:14.
18. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med*. 1996;15(24):2733-2749.
19. Ioannidis JP. Ranking antidepressants. *Lancet*. 2009;373(9677):1759-1760, author reply 1761-1762.
20. Gartlehner G, Hansen RA, Morgan LC, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. *Ann Intern Med*. 2011;155(11):772-785.

21. Nixon RM, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med.* 2007;26(6):1237-1254.
22. Mills EJ, Wu P, Chong G, et al. Efficacy and safety of statin treatment for cardiovascular disease: a network meta-analysis of 170,255 patients from 76 randomized trials. *QJM*. 2011;104(2):109-124.
23. Mills EJ, Wu P, Lockhart I, Thorlund K, Puhan M, Ebbert JO. Comparisons of high-dose and combination nicotine replacement therapy, varenicline, and bupropion for smoking cessation: a systematic review and multiple treatment meta-analysis. *Ann Med.* 2012;44(6):588-597.
24. Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med*. 2008;148(7):544-553.
25. Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. *PLoS One*. 2011;6(1):e16237.
26. Davey-Smith G, Egger MG. Going beyond the grand mean: subgroup analysis in meta-analysis of randomised trials. In: *Systematic Reviews in Health Care: Meta-analysis in context.* 2nd ed. London, England: BMJ Publishing Group; 2001:143-156.
27. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-1573.
28. Jansen J, Schmid C, Salanti G. When do indirect and mixed treatment comparisons result in invalid findings? A graphical explanation. 19th Cochrane Colloquium Madrid, Spain October 19-22, 2011. 2011:P3B379.
29. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *J Clin Epidemiol*. 2008;61(5):455-463.
30. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101(474):447-459.
31. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7-8):932-944.
32. Zhang WY, Li Wan Po A. Analgesic efficacy of paracetamol and its combination with codeine and caffeine in surgical pain—a meta-analysis. *J Clin Pharm Ther*. 1996;21(4):261-282.
33. 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(2):163-171.
34. Golfinopoulos V, Salanti G, Pavlidis N, Ioannidis JP. Survival and disease-progression benefits with treatment regimens for advanced colorectal cancer: a meta-analysis. *Lancet Oncol*. 2007;8(10):898-911.
35. Diels J, Cure S, Gavart S. The comparative efficacy of telaprevir versus boceprevir in treatment-naive and treatment experienced patients with genotype 1 chronic hepatitis C virus infection: a mixed treatment comparison analysis. Paper presented at: 14th Annual International Society for Pharmaceutical Outcomes Research (ISPOR) European Congress; November 5-8, 2011; Madrid, Spain.

36. Diels J, Cure S, Gavart S. The comparative efficacy of telaprevir versus boceprevir in treatment-naive and treatment-experienced patients with genotype 1 chronic hepatitis. *Value Health*. 2011;14(7):A266.
37. Higgins JP, Green S. Analysing data and undertaking meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Oxford: Wiley & Sons; 2008.
38. Aaron SD, Fergusson D, Marks GB, et al; Canadian Thoracic Society/Canadian Respiratory Clinical Research Consortium. Counting, analysing and reporting exacerbations of COPD in randomised controlled trials. *Thorax*. 2008;63(2):122-128.
39. Mills EJ, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol*. 2011;3:107-129.
40. Sciarretta S, Palano F, Tocci G, Baldini R, Volpe M. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. *Arch Intern Med*. 2011;171(5):384-394.
41. Psaty BM, Lumley T, Furberg CD, et al. Health outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. *JAMA*. 2003;289(19):2534-2544.
42. Hernandez AV, Walker E, Ioannidis JP, Kattan MW. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. *Am Heart J*. 2008;156(1):23-30.
43. Ioannidis JP, Evans SJ, Gøtzsche PC, et al; CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141(10):781-788.
44. Bajwa Z, Sabahat A. Acute treatment of migraine in adults. UpToDate website. http://www.uptodate.com/contents/acute-treatment-of-migraine-in-adults. Accessed August 4, 2014.
45. Ford AC, Talley NJ, Spiegel BM, et al. Effect of fibre, antispasmodics, and peppermint oil in the treatment of irritable bowel syndrome: systematic review and meta-analysis. *BMJ*. 2008;337:a2313.
46. Schmitz S, Adams R, Walsh CD, Barry M, FitzGerald O. A mixed treatment comparison of the efficacy of anti-TNF agents in rheumatoid arthritis for methotrexate non-responders demonstrates differences between treatments: a Bayesian approach. *Ann Rheum Dis*. 2012;71(2):225-230.
47. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010;340:c117.
48. 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.