ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1303–1310

Journal of Clinical Epidemiology

GRADE guidelines: 8. Rating the quality of evidence—indirectness

Gordon H. Guyatt[a,b,*], Andrew D. Oxman[c], Regina Kunz[d,e], James Woodcock[f], Jan Brozek[a], Mark Helfand[g], Pablo Alonso-Coello[h], Yngve Falck-Ytter[i,j], Roman Jaeschke[b], Gunn Vist[c], Elie A. Akl[k], Piet N. Post[l], Susan Norris[m], Joerg Meerpohl[n,o], Vijay K. Shukla[p], Mona Nasser[q], Holger J. Schünemann[a,b], The GRADE Working Group[1]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway*
[d]*Academy of Swiss Insurance Medicine (asim), University Hospital Basel Petergraben 4, CH-4031, Basel, Switzerland*
[e]*Basel Institute of Clinical Epidemiology, University Hospital Basel Hebelstrasse 10, 4031 Basel, Switzerland*
[f]*London School of Hygiene and Tropical Medicine, London, United Kingdom*
[g]*Oregon Evidence-Based Practice Center, Oregon Health and Science University, Portland VA Medical Center, Portland, OR, USA*
[h]*Iberoamerican Cochrane Center-Servicio de Epidemiología Clínica y Salud Pública and CIBER de Epidemiología y Salud Pública (CIBERESP), Hospital de Sant Pau, Universidad Autónoma de Barcelona, Barcelona 08041, Spain*
[i]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[j]*University of Oxford, Oxford, United Kingdom*
[k]*Department of Medicine, State University of New York at Buffalo, NY, USA*
[l]*Dutch Institute for Healthcare Improvement CBO, Utrecht, The Netherlands*
[m]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[n]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[o]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*
[p]*Canadian Agency for Drugs and Technology in Health (CADTH), Ottawa K1S 5S8, Canada*
[q]*Department of health information, Institute for Quality and Efficiency in Health care (IQWiG), Cologne, Germany*

Accepted 18 April 2011; Published online 30 July 2011

Abstract

Direct evidence comes from research that directly compares the interventions in which we are interested when applied to the populations in which we are interested and measures outcomes important to patients. Evidence can be indirect in one of four ways. First, patients may differ from those of interest (the term applicability is often used for this form of indirectness). Secondly, the intervention tested may differ from the intervention of interest. Decisions regarding indirectness of patients and interventions depend on an understanding of whether biological or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect.

Thirdly, outcomes may differ from those of primary interest—for instance, surrogate outcomes that are not themselves important, but measured in the presumption that changes in the surrogate reflect changes in an outcome important to patients.

A fourth type of indirectness, conceptually different from the first three, occurs when clinicians must choose between interventions that have not been tested in head-to-head comparisons. Making comparisons between treatments under these circumstances requires specific statistical methods and will be rated down in quality one or two levels depending on the extent of differences between the patient populations, co-interventions, measurements of the outcome, and the methods of the trials of the candidate interventions.

Keywords: GRADE; Quality of evidence; Indirectness; Indirect comparisons; Applicability; Generalizability

[1] The GRADE system has been developed by the GRADE Working Group. The named authors drafted and revised this article. A complete list of contributors to this series can be found on the journal's Web site at www.elsevier.com.

* Corresponding author. CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada. Tel.: 905-527-4322; fax: 905-523-8781.

E-mail address: guyatt@mcmaster.ca (G.H. Guyatt).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.04.014

Key points

- Quality of evidence (our confidence in estimates of effect) may decrease when substantial differences exist between the population, the intervention, or the outcomes measured in relevant research studies and those under consideration in a guideline or systematic review.
- Quality of evidence decreases if head-to-head comparisons are unavailable. Such instances require falling back on indirect comparisons in which, for example, we make inferences about the relative effect of two interventions on the basis of their comparison not with one another, but with a third or control condition.

1. Introduction

Previous articles in this series presenting GRADE's approach to systematic reviews and clinical guidelines have dealt with framing the question, defined quality of evidence, and described GRADE's approach to rating down the quality of a body of evidence because of problems with bias, imprecision, and inconsistency. In this article, we deal with another potential problem: indirectness.

2. Four types of indirectness

We are more confident in the results when we have direct evidence. By direct evidence, we mean research that directly compares the interventions in which we are interested delivered to the populations in which we are interested and measures the outcomes important to patients. Thus, we can have concerns about indirectness when the population, intervention, or outcomes differ from those in which we are interested (Table 1). A fourth, different type of indirectness, occurs when there are no head-to-head comparisons between the alternative management strategies under comparison (Table 1). Indirectness of outcomes and indirect comparisons are equally relevant to systematic reviews and practice guidelines; indirectness related to populations and interventions (sometimes referred to as applicability) is more relevant to guidelines.

2.1. Indirectness: differences in population (applicability)

The first type of indirectness includes differences between the population of interest and those who have participated in relevant studies. Systematic reviews will include only patients who meet the reviews' eligibility criteria; thus, in a sense, evidence regarding patients is direct by definition.

There may, however, be exceptions. For example, a systematic review might have an a priori hypothesis that a drug would have different effects in children than in adults based on what is known about the mechanism of action. If no studies were found that tested the drug in children, the review authors might conclude that the effects in children were less certain than in adults, based on the indirectness of the evidence for children.

Differences between the population of interest and those in studies are a common problem for guideline developers who will seek the best evidence relevant to their question. For instance, a World Health Organization guideline panel addressed the treatment of infection with avian influenza A virus but needed to use evidence from seasonal influenza (Table 1; Box 1) [1].

Less extreme differences in patients (or the conditions from which they suffer) would lead to rating down only one level, or even no rating down whatsoever. Because randomized trial eligibility criteria often exclude patients with comorbidity, as guideline developers begin to address issues of multiple coexisting conditions (patients with, for instance, heart failure and asthma) they will often need to consider issues of indirectness. Some population differences may be partly addressed by subgroup analyses within the trials or reviews that check the robustness of the results across population factors such as age, gender, or disease severity. For example, pooled analyses of large-scale trials of statins show highly consistent relative risk (RR) reductions across a wide variety of subpopulations.

In general, one should not rate down for population differences unless one has compelling reason to think that the biology in the population of interest is so different from that

Table 1. Evidence is lower quality if comparisons are indirect

Question of interest	Source of indirectness
Oseltamivir for prophylaxis of avian flu caused by influenza A virus	*Differences in population*: Randomized trials of oseltamivir are available for seasonal influenza, but not for avian flu
Colonoscopic screening for prevention of colon cancer mortality	*Differences in intervention*: Randomized trials of fecal occult blood screening provide indirect evidence bearing on the potential effectiveness of colonoscopy
Sevelamer- vs. calcium-based phosphate binders in chronic renal failure	*Differences in outcome*: Reducing the calcium-phosphate load is hypothesized to reduce vascular calcification, which is hypothesized to reduce vascular events
Choice of antidepressant	*Indirect comparison*: Some antidepressants have been compared directly with others, but many have not

Box 1 Indirectness of population: avian influenza

High-quality randomized trials have demonstrated the effectiveness of antiviral treatment for seasonal influenza. The panel judged that the biology of seasonal influenza was sufficiently different from that of avian influenza (that is, the avian influenza organism may be far less responsive to the available antiviral agents than seasonal influenza) that the evidence required rating down by two levels for indirectness.

of the population tested that the magnitude of effect will differ substantially. Most often, this will not be the case. Note that we are referring here to consistency in RR: differences in baseline risk or control event rate in subpopulations will, on many occasions, lead to difference in absolute effect between subgroups.

The above discussion refers to different human populations, but sometimes the only evidence will be from animal studies, such as rats or primates. In general, we would rate such evidence down two levels for indirectness. Animal studies may, however, provide an important indication of drug toxicity. Although toxicity data from animals does not reliably predict toxicity in humans, evidence of animal toxicity should engender caution in recommendations.

Another type of nonhuman study may generate high-quality evidence. Consider laboratory evidence of change in resistance patterns of bacteria to antimicrobial agents (e.g., the emergence of methicillin-resistant staphylococcus aureus—MRSA). These laboratory findings may constitute high-quality evidence for the superiority of antibiotics to which MRSA is sensitive vs. methicillin as the initial treatment of suspected staphylococcus sepsis in settings in which MRSA is highly prevalent.

2.2. Indirectness: differences in interventions (applicability)

As for the population, systematic reviewers will clearly specify the interventions of interest in their eligibility criteria, ensuring that only directly relevant studies will be eligible. Again, however, there may be exceptions. For example, a systematic review might have an a priori hypothesis that a surgical procedure would have different effects when undertaken by subspecialists in referral centers compared with general surgeons in the community. If they found no studies that tested the procedure in community hospitals, review authors might conclude that the effects of the procedure undertaken by community general surgeons were uncertain.

Guideline developers may often find the best evidence addressing their question in trials of related interventions. For example, a guideline addressing the value of colonoscopic screening for colon cancer will find the randomized control trials (RCTs) of fecal occult blood screening that showed a decrease in colon cancer mortality in people receiving the intervention of relevance. Whether to rate down one or two levels in this context is a matter of judgment.

There may be instances in which the intervention differs, but authors may conclude that there is no need to rate down for quality. For example, older trials show a high efficacy of intramuscular penicillin for gonococcal infection, but guidelines might reasonably recommend alternative antibiotic regimes based on current local in vitro resistance patterns, and consider the evidence as high quality.

Interventions may be delivered differently in different settings. For instance, a systematic review of music therapies for autism found that trials tested structured approaches that are used more commonly in North America than in Europe. Because the interventions differ, the results from structured approaches are more applicable to North America and the results of less structured approaches are more applicable in Europe. Issues of setting are particularly crucial for the outcome of resource use (cost). The resources required (or at least used) for a particular intervention may vary widely across settings, and the opportunity cost (what alternatives could be purchased for the same money) differs to an even greater extent.

Guideline panelists should consider rating down the quality of the evidence if the intervention cannot be implemented with the same rigor or technical sophistication in their setting as in the RCTs from which the data come. Carotid endarterectomy provides a commonly cited example of such a situation [2]. Indirectness of this sort becomes a major issue—particularly for lower-income countries—for resource-intensive interventions. We have noted this issue under ''setting'' for indirectness of interventions, in which we referred to how music therapy for autism may be delivered differently in one jurisdiction than another. The same is true for other complex interventions such as rehabilitation programs and public health interventions. There may be important differences in implementation across settings that can weaken inferences regarding applicability.

As with all other aspects of rating quality of evidence, there is a continuum of similarity of the intervention that will require judgment. It is rare, and usually unnecessary, for the intended populations and interventions to be identical to those in the studies, and we should only rate down if the differences are considered sufficient to make a difference in outcome likely. For example, trials of simvastatin show cardiovascular mortality reductions: suggesting night rather than morning dosing (because of greater cholesterol reduction) would not warrant rating down for differences in the intervention. A new statin with available evidence only from lipid levels might, however, require rating down quality for indirectness, and trials of a new class of cholesterol-lowering agents in which RCTs have not addressed impact on cardiovascular events would certainly require rating down for indirectness. One could conceptualize this as

rating down for either indirectness of interventions or indirectness of outcomes.

2.3. Indirectness: differences in outcome measures (surrogate outcomes)

GRADE specifies that both those conducting systematic reviews and those developing practice guidelines should begin by specifying every important outcome of interest. The available studies may have measured the impact of the intervention of interest on outcomes related to, but different from, those of primary importance to patients.

The difference between desired and measured outcomes may relate to time frame. For example, a systematic review of behavioral and cognitive-behavioral interventions for outwardly directed aggressive behavior in people with learning disabilities showed that a program of 3-week relaxation training significantly reduced disruptive behaviors at 3 months [3]. Unfortunately, no eligible trial assessed the review authors' predefined outcome of interest, the long-term impact defined as effect at 9 months or greater. The argument for rating down becomes even stronger when one considers that other types of behavioral interventions have shown an early beneficial effect that was not sustained at 6 months follow-up [3]. When there is a discrepancy between the time frame of measurement and that of interest, whether to rate down by one or two levels will depend on the magnitude of the discrepancy. In this case, one could argue for either option.

Another source of indirectness related to measurement of outcomes is the use of substitute or surrogate endpoints in place of the patient-important outcome of interest. Table 2 lists a number of such surrogate measures that are common in current clinical investigation.

Table 3 presents the logic of patient-important and surrogate outcomes as applied to disturbances in calcium and phosphate metabolism in patients with end-stage renal disease. Hyperphosphatemia is associated with abnormal bone fragility and consequent fractures; soft tissue calcification and associated pain; coronary calcification and associated myocardial infarction; and possible increased mortality. These adverse outcomes are the important endpoints in treating the calcium/phosphate abnormalities.

Up to now, however, RCTs of alternative therapeutic interventions have focused on measures of calcium/phosphate metabolism. In general, the use of a surrogate outcome requires rating down the quality of evidence by one, or even two, levels. Consideration of the biology, mechanism, and natural history of the disease can be helpful in making a decision about indirectness. For instance, because concentrations of calcium and phosphate are far removed in the putative causal pathway from the patient-important endpoints, we would rate down the quality of evidence with respect to this outcome by two levels (Table 3). Surrogates that are closer in the putative causal pathway to the adverse outcomes are coronary calcification (for myocardial infarction), bone density (for fractures), and soft-tissue calcification (for pain), and these outcomes warrant rating down by only one level for indirectness.

A systematic review suggesting a benefit of low molecular weight heparin vs. unfractionated heparin for perioperative thromboprophylaxis in patients with cancer provides an example in which rating down by just one level for indirectness is probably appropriate. The confidence intervals (CIs) around reduction in the important outcome, symptomatic deep venous thrombosis (DVT), were extremely wide (RR 0.73; 95% CI: 0.23, 2.28). When the outcome included the surrogate, asymptomatic DVT (which provided most events), the difference in favor of low molecular weight heparin was much more convincing (RR = 0.72; 95% CI: 0.55, 0.94) [4]. Convincing evidence of reduction in asymptomatic events provides, in our view, moderate quality evidence of a reduction in symptomatic events.

Rarely, surrogates are sufficiently well established that review authors or guideline panelists should choose not to rate down quality of evidence for indirectness. In our view, this should be restricted to situations in which, within the same class of drug (e.g., beta-blockers, calcium antagonists, diuretics, bisphosphonates), changes in the surrogate have repeatedly proved closely related to changes in the patient-important outcome in the context of RCTs. One might use this rationale, for example, to justify not rating

Table 2. Examples of surrogate outcomes

Condition	Patient-important outcome(s)	Surrogate outcome(s)
Diabetes mellitus	Diabetic symptoms, hospital admission, complications (cardiovascular, eye, renal, neuropathic)	Blood glucose, A1C
Hypertension	Cardiovascular death, myocardial infarction, stroke	Blood pressure
Dementia	Patient function, behavior, caregiver burden	Cognitive function
Osteoporosis	Fractures	Bone density
Adult Respiratory Distress Syndrome	Mortality	Oxygenation
End-stage renal disease	Quality of life, morbidity (such as shunt thrombosis or heart failure), mortality	Hemoglobin
Venous thrombosis	Symptomatic venous thrombosis	Asymptomatic venous thrombosis
Chronic respiratory disease	Quality of life, exacerbations, mortality	Pulmonary function, exercise capacity
Cardiovascular disease/risk	Vascular events, mortality	Serum lipids

Table 3. Surrogate and patient-important outcomes for phosphate lowering drugs in patients with renal failure and hyperphosphatemia

	Surrogate outcomes	
Patient-important outcomes	***Indirect*** **(Lower the quality of evidence by one level)**	***Very indirect*** **(Lower the quality of evidence by two levels)**
Myocardial infarction	Coronary calcification	Measures of calcium/phosphate metabolism
Fractures	Bone density	
Pain because of soft-tissue calcification	Soft-tissue calcification	

down low-density lipoprotein (LDL) as a surrogate for coronary events in evaluating the evidence from RCTs of a new statin. One would, however, rate down for indirect outcomes the evidence from RCTs of another class of cholesterol-lowering agents (e.g., ezetimibe) if the outcome measure was LDL rather than coronary events. Even this highly restricted criterion for not rating down a surrogate (multiple randomized trials within a single drug class show a clear and consistent relationship between change in the surrogate and an effect measure such as RR reduction) may be problematic (Box 2).

Investigators may use sophisticated statistical approaches to examine the relationship between a surrogate and a patient-important outcome. For instance, investigators examined the ''validity'' of progression-free survival as a surrogate for overall survival for anthracycline- and taxine-based chemotherapy for advanced breast cancer [5]. They found a statistically significant association between progression-free and overall survival in the randomized trials they analyzed, but predicting overall survival using progression-free survival remained fraught with uncertainty. Rating down quality by one level for the surrogate would be appropriate in this situation.

Several groups have developed systems for rating the ''validity'' of a surrogate [6,7,16]. Each of these systems finds evidence from surrogates convincing only when the association has been strongly and repeatedly established in RCTs. Systematic review authors and guideline developers may wish to refer to these systems when pondering whether to rate down for indirectness of outcomes.

Box 2 Arguments against ever-considering evidence from surrogates of high quality

One might well be tempted to assume a new statin that improves lipid profiles in the same way as older statins would result in similar improvement in cardiovascular risk. Authorities have, however, raised arguments about assuming that low-density lipoprotein (LDL) reductions with a new statin will translate into the expected reduction in cardiovascular risk [13,14]. Indeed, in one large trial in hemodialysis patients, large reductions in LDL failed to effect reductions in cardiovascular events [15]. In addition, deciding what constitutes a class of drugs (e.g., all beta-blockers; all cardioselective beta-lockers; all beta-blockers with or without alpha-blocking properties) is not straightforward [16,17]. Finally, from a clinical point of view, even if one accepts that a surrogate provides high-quality evidence regarding benefit, a new agent may have a different—and highly problematic—side effect profile. Note, for instance, cerivastatin's greatly increased—relative to other statins—propensity to cause life-threatening rhabdomyolysis.

2.4. *Indirectness: indirect comparisons*

The final type of indirectness occurs when we have no direct (i.e., head-to-head) comparisons between two or more interventions of interest. For instance, consider a comparison of two active drugs, A and B. Although RCTs comparing A and B may be unavailable, RCTs may have compared A to placebo and B to placebo. Such trials allow indirect comparisons of the magnitude of effect of A and B. Such evidence is of lower quality than head-to-head comparisons of A and B would provide.

Indirect comparisons of prophylactic treatments for osteoporotic fractures illustrate the challenges of indirect comparisons. Trials of different agents suggest apparent differences in RR reduction, tempting one to attribute these differences to varying effectiveness of the drugs under consideration. The trials, however, enrolled different groups of patients; some may be more responsive than others. In addition, trials varied in criteria for diagnosis of both vertebral and nonvertebral fractures. It may be these differences, rather than differences in the effectiveness of the interventions, that are responsible for variation in RR [8]. A systematic review of different doses of aspirin illustrates the difficulties of inferences from indirect comparisons (Box 3).

The validity of the indirect comparison rests on the assumption that factors in the design of the trial (the patients, co-interventions, measurement of outcomes) and the methodological quality are not sufficiently different to result in different effects (in other words, true differences in effect explain all apparent differences). Some authors refer to this as the ''similarity assumption'' [9]. Because this assumption is always in some doubt, indirect comparisons always warrant rating down by one level in quality of evidence. Whether to rate down two levels depends on the plausibility that alternative factors (population, interventions, co-interventions, outcomes, and study methods) explain or obscure differences in effect. Of the many

Box 3 Difficulties making inferences from indirect comparisons: low- vs. medium-dose aspirin

A systematic review considered the relative merits of low dose (50−150 mg daily) vs. medium dose (300−325 mg daily) of aspirin to prevent graft occlusion after coronary artery bypass surgery [18]. Authors found five relevant trials that compared aspirin with placebo, of which two tested medium dose and three low-dose aspirin. The pooled relative risk (RR) of the likelihood of a graft occlusion was 0.74 (95% confidence interval [CI]: 0.60, 0.91) in the low-dose trial and 0.55 (95% CI: 0.28, 0.82) in the medium-dose trials. The RR of medium vs. low dose was 0.74 (95% CI: 0.52, 1.06; $P = 0.10$) suggesting (but not very convincingly) the possibility of a larger effect with the medium-dose regimens.

This comparison is weaker than if the randomized trials had compared the two aspirin dose regimens directly because there are other study characteristics that might be responsible for any differences found (or resulted in undetected differences that in fact exist). Compared with the low dose vs. placebo trials, in medium dose vs. placebo trials, the patients studied may be different, effective or harmful interventions other than the therapy under investigation may have been differently administered, and outcomes may have been measured differently (e.g., dissimilar criteria for events or varying duration of follow-up). Differences in study methods may also explain the results: trials with a higher risk of bias may result in smaller—or more likely larger—treatment effects.

challenging judgments that rating quality of evidence demands, this is one of the most difficult.

The judgment is made more difficult yet by the necessity to consider the statistical approaches that investigators have taken in making indirect comparisons. Simply using the results from the active arms in two or more studies is naïve and potentially misleading. More sophisticated statistical approaches that consider differences between active and control arms are more appropriate [10,11].

The comparison of low- vs. medium-dose aspirin regimens (Box 3) used a valid statistical approach to compare the RRs in one set of trials to the RR in the other set. The review authors present data suggesting that the trials enrolled patients who were very similar with respect to mean age (56−60 years), sex distribution (83−100% men), proportion of smokers (65−68% in the two trials reporting), proportion of hypertensive patients (31−53% in the four trials reporting), and mean cholesterol (5.7−7.2 mmol/L). The authors did not mention whether the two sets of studies differed in the use of a cointervention—aggressiveness of antihypertensive treatment or the use of lipid lowering agents, for instance. In terms of methods, one trial in each set standardized surgical procedures, all were blinded and included a placebo arm, two medium-dose and one low-dose trial reported formal randomization by research-coordinating centers or pharmacy, and one trial in each group reported independent angiographic assessment of vein graft patency. Both sets of trials had very high loss to follow-up (i.e., no angiography)—three of five trials reported rates of more than 50%.

On balance, we would rate down the quality of the evidence only one level for indirectness. The decision in this case has little effect on clinical decision making in that other considerations (risk of bias—high loss to follow-up, imprecision—wide CIs around the RR in moderate vs. low-dose trials, and indirectness of outcomes—graft occlusion is a surrogate for events such as myocardial infarction and cardiovascular deaths) already place this as low-quality evidence. The indirect comparison leaves us with very low-quality evidence.

Increasingly, recommendations must simultaneously address more than two interventions. For instance, possible approaches to thrombolysis in myocardial infarction include streptokinase, alteplase, reteplase, and tenecteplase [12]. Attempts to address such issues of the relative effectiveness of multiple interventions inevitably involve indirect comparisons. These meta-analyses have received different labels; currently popular terms include ‘‘network meta-analyses,’’ ‘‘mixed treatment comparison,’’ and ‘‘multiple treatments meta-analysis.’’

There are both simple, inappropriate approaches, and a number of sophisticated appropriate statistical approaches available for assessing simultaneous multiple comparisons. A variety of recently developed Bayesian statistical methods may help in generating estimates of the relative effectiveness of multiple interventions, but these methods may give different estimates. This raises the possibility of bias, and the issue of the best-quality indirect analysis is unsettled. Their confident application requires, in addition to indirect comparison evidence, substantial evidence from direct comparisons—evidence that is often unavailable [12]. Ascertaining the extent to which patients, co-interventions, measurement of outcomes, and risk of bias in studies of multiple interventions are similar presents another major challenge. Interpretation when direct and indirect evidence is inconsistent is uncertain, and may warrant rating down the direct evidence for inconsistency. A recent simultaneous treatment comparison illustrates the challenges of evaluating such studies (Box 4). The methods to conduct and assess such network meta-analyses, including GRADE’s approach, remain in evolution. The coming years should see refinement in criteria for judging the quality of evidence from network meta-analyses.

A final point is that it is possible, at least in theory, for indirect comparisons to yield more accurate results than direct comparisons. This could be true if direct comparisons suffer from risk of bias that indirect comparisons do not. This may occur if the direct comparisons are conducted by those with an investment in the result (e.g., industry).

Box 4 An example of the challenges of network meta-analysis

Investigators conducted a simultaneous treatment comparison of 12 new generation antidepressants [19]. The authors evaluated 117 randomized trials involving over 25,000 patients; their article provides no information about the similarity of the patients (other than that they all had major unipolar depression), or about cointervention (behavioral therapies, for instance). In correspondence with the authors, however, they indicated that they excluded trials with treatment-resistant depression, argued that different types of depression have similar treatment responses, and that it is very likely that patients did not receive important cointervention. With respect to risk of bias, the authors tell us, using the Cochrane collaboration approach to assessing risk of bias [20] that risk of bias in most studies was ''unclear,'' and 12 were at low risk of bias; presumably a small number was at high risk of bias. This is helpful, although ''unclear'' represents a wide range of risk of bias.

All studies involved head-to-head comparisons between at least two of the 12 drugs; the 117 trials involved 70 individual comparisons (e.g., two comparisons between fluoxetine and fluvoxamine). The authors reported statistically significant differences between direct and indirect comparisons in only three of 70 comparisons of drug response. The power of such tests was, however, not likely high. Overall, we would be inclined to take a cautious approach to this network meta-analysis and rate down two levels for indirectness.

3. Mechanism

Another type of indirect evidence that we have not addressed relates to mechanism of action. The GRADE system does not rate evidence either up or down based on the mechanism or pathophysiological basis of a treatment. RCTs typically begin with a reasonable expectation of success based, to some degree, on biological rationale. But judgments of exactly how strong is the rationale are easily open to dispute, and GRADE does not suggest using them directly as a basis for rating evidence quality up or down.

Mechanism does, however, have multiple roles in the evaluation of evidence: in selecting studies for systematic reviews, in the applicability of evidence to different interventions or populations, in judging whether to believe subgroup analyses, and in deciding the extent to which one rates down quality of evidence based on surrogate outcomes. Although it would make little sense to pool studies based on similar costs or color of tablets, treatments with similar mechanism are commonly meta-analyzed. Because no two studies will have exactly the same eligibility criteria and interventions, judgments based on our biological understanding are necessary to determine which studies to include in generating a single pooled estimate of effect.

Similarly, we need to make judgments based on mechanism to apply evidence about treatments. For example, if a trial that included patients aged 50−70 years showed effect, then we would undoubtedly be happy to apply the results to 49- or 71-year olds (and likely well younger than 49 and well older than 71 years) but not to children. If a study showed 5 days of antibiotics were effective, then we might be happy to use 7 days but not 3 days.

Judgments regarding surrogate outcomes may, however, be more complex. For example, consider a three-dose vaccine that reduced the incidence of the target illness. We might be happy to consider an accelerated delivery of three doses of exactly the same vaccine to be as effective as the original if the studies showed that the standard and accelerated three-dose regimes had a similar serological response (i.e., we might not rate down quality of evidence because of the surrogate outcome of serological response). However, we might rate down for use of a surrogate outcome if a new class of antihypertensive agents (e.g., the direct renin inhibitor aliskiren, recently licensed in the United States) showed a similar reduction in blood pressure to existing agents but without evidence of benefit on patient-important outcomes.

4. Simultaneous consideration of all types of indirectness

Guideline developers will usually need to consider the combined effect of all the four types of indirectness—and problems in more than one may suggest the need to rate down two levels in the quality of evidence. This consideration is not a simple additive process, but rather a judgment about whether any, and how much, rating down is warranted. In general, evidence based on surrogate outcomes should usually trigger rating down, whereas the other types of indirectness will require a more considered judgment.

References

[1] Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis 2007;7(1):21−31.

[2] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365(9453):82−93.

[3] Hassiotis A, Hall I. Behavioural and cognitive-behavioural interventions for outwardly-directed aggressive behaviour in people with learning disabilities. Cochrane Database Syst Rev 2004;1: CD003406. DOI:10.1002/14651858.CD003406.pub2.

[4] Akl EA, Terrenato I, Barba M, Sperati F, Sempos EV, Muti P, et al. Low-molecular-weight heparin vs unfractionated heparin for

perioperative thromboprophylaxis in patients with cancer: a systematic review and meta-analysis. Arch Intern Med 2008;168:1261−9.

[5] Miksad RA, Zietemann V, Gothe R, Schwarzer R, Conrads-Frank A, Schnell-Inderst P, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. Int J Technol Assess Health Care 2008;24(4):371−83.

[6] Lassere MN, Johnson KR, Boers M, Tugwell P, Brooks P, Simon L, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol 2007;34: 607−15.

[7] Australian Government Department of Health and Ageing. Report of the Surrogate to Final Outcome Working Group to the Pharmaceutical Benefits Advisory Committee: a framework for evaluating proposed surrogate measures and their use in submissions to PBAC. 2009.

[8] Sebba A. Comparing non-vertebral fracture risk reduction with osteoporosis therapies: looking beneath the surface. Osteoporos Int 2009;20:675−86.

[9] Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009;338:b1147.

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

[11] Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9(26):1−134, iii−iv.

[12] Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331(7521):897−900.

[13] de Lorenzo F, Feher M, Martin J, Collot-Teixeira S, Dotsenko O, McGregor JL. Statin therapy-evidence beyond lipid lowering contributing to plaque stability. Curr Med Chem 2006;13:3385−93.

[14] Fisman EZ, Adler Y, Tenenbaum A. Statins research unfinished saga: desirability versus feasibility. Cardiovasc Diabetol 2005;4(1):8.

[15] Wanner C, Krane V, März W, Olschewski M, Mann JF, Ruf G, et al. Atorvastatin in patients with type 2 diabetes mellitus undergoing hemodialysis. N Engl J Med 2005;353(3):238−48.

[16] Bucher H, Kunz R, Cook D, Holbrook A, Guyatt G. Surrogate outcomes. In: Guyatt G, Rennie D, Meade M, Cook D, editors. The users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.

[17] Kunz R, Bucher H, McAlister F, Holbrook A, Guyatt G. Drug class effects. In: Guyatt G, Rennie D, Meade M, Cook D, editors. The users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.

[18] Lim E, Ali Z, Ali A, Routledge T, Edmonds L, Altman DG, et al. Indirect comparison meta-analysis of aspirin therapy after coronary surgery. BMJ 2003;327(7427):1309.

[19] 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(9665):746−58.

[20] Higgins JP, Altman D. Assessing the risk of bias in included studies. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions 5.0.1. Chichester, UK: John Wiley & Sons; 2008.