ICER Evidence Rating Matrix:

A User's Guide

Dan Ollendorf, MPH
Chief Review Officer

Steven D. Pearson, MD, MSc, FRCP
President

ICER Evidence Rating Matrix User Guide Brought to you by the CER Collaborative:

A Collaboration of the Academy of Managed Care Pharmacy, the International Society for Pharmacoeconomics and Outcomes Research and the National Pharmaceutical Council

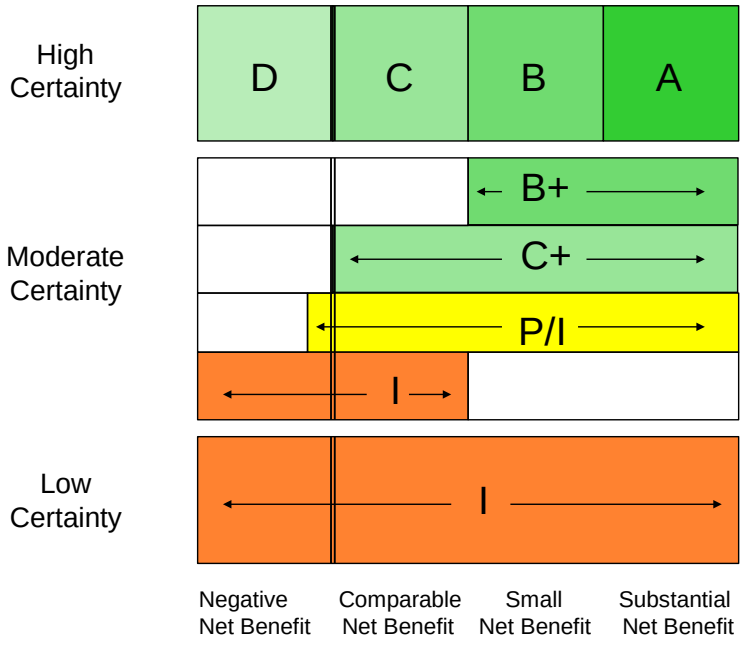
Table of Contents

Executive Summary: Using the Matrix

Formulary decisions require a rigorous evaluation of available evidence, a process that entails judgments regarding the quality of individual clinical studies and, ultimately, an assessment of the entire body of evidence regarding a therapeutic agent. To support this latter step, the Institute for Clinical and Economic Review (ICER) has developed the ICER Evidence Rating Matrix™. This user's guide to the ICER Matrix was developed with funding provided by the Comparative Effectiveness Research Collaborative Initiative (CER-CI), a joint initiative of the Academy of Managed Care Pharmacy, the International Society for Pharmacoeconomics and Outcomes Research, and the National Pharmaceutical Council (www.cercollaborative.org). The ICER Matrix presents a framework for evaluating the comparative benefits and risks of therapies in a consistent, transparent system leading to an evidence rating that can guide coverage and formulary placement decisions. The purpose of this user's guide is to help members of Pharmacy and Therapeutics Committees and other decision-makers understand the approach embodied in the matrix, and to help them apply it in a reliable, consistent fashion.

The updated ICER Evidence Rating Matrix is shown below, with a key to the single letter ratings on the following page. Fundamentally, the evidence rating reflects a joint judgment of two critical components:

a) The **magnitude** of the difference between a therapeutic agent and its comparator in "net health benefit" – the balance between clinical benefits and risks and/or adverse effects (horizontal axis); AND
b) The level of **certainty** that you have in your best point estimate of net health benefit (vertical axis).



The letter ratings are listed below, according to the level of certainty in the best estimate of net health benefit. They are described in further detail on pages 5-6.

High Certainty

A = Superior
B = Incremental
C = Comparable
D = Inferior

Moderate Certainty

B+=Incremental or Better
C+=Comparable or Better
P/I = Promising but Inconclusive
I = Insufficient

Low Certainty

I = Insufficient

Steps in Applying the ICER Evidence Rating Matrix

1. **Establish the specific focus of the comparison to be made and the scope of evidence you will be considering.** This process is sometimes referred to as determining the "PICO" – the **P**opulation, **I**ntervention, **C**omparator(s), and **O**utcomes of interest. Depending on the comparison, it is often helpful to also define the specific **T**ime Horizon and **S**etting that will be considered relevant.

2. **Estimate the magnitude of the comparative net health benefit.** Working from the scope of evidence established, it is important to quantify findings from the body of evidence on specific clinical benefits, risks, and other potentially important outcomes, such as adherence, so you can compare these side-by-side for the therapeutic agent and comparator. Some organizations compare each outcome, risk, etc. separately without using a quantitative measure to try to sum the overall comparative balance of benefits and risks between the therapeutic agent and the comparator. For these organizations the estimate of comparative net health benefit must be made qualitatively. Other organizations summarize the balance of benefits and risks using formal mathematical approaches such as health utility analysis, which generates a quantitative summary measure known as the quality-adjusted life year (QALY). What is most important, however, is full and transparent documentation of your rationale for assigning the magnitude of comparative net health benefit into one of four possible categories:

    - ***Negative:*** the drug produces a net health benefit inferior to that of the comparator
    - ***Comparable:*** the drug produces a net health benefit comparable to that of the comparator
    - ***Small:*** the drug produces a small positive net health benefit relative to the comparator
    - ***Substantial:*** the drug produces a substantial (moderate-large) positive net health benefit relative to the comparator

3. **Assign a level of certainty to the estimate of comparative net health benefit.** Given the strength of the evidence on comparative benefits and risks, a "conceptual confidence interval" around the original estimate of comparative net health benefit can be made, leading you to an assignment of the overall level of certainty in that estimate. Rather than assigning certainty by using a fixed equation weighting different attributes of the body of evidence, we recommend formal documentation of the consideration of 5 major domains related to strength of evidence: (1) Level of Bias—how much risk of bias is there in the study designs that comprise the entire evidence base? (2) Applicability—how generalizable are the results to real-world populations and conditions? (3) Consistency—do the studies produce similar treatment effects, or do they conflict in some ways? (4) Directness—are direct or indirect comparisons of therapies available, and/or are direct patient outcomes measured or only surrogate outcomes, and if surrogate outcomes only, how validated are these measures? (5) Precision—does the overall database include enough robust data to provide precise estimates of benefits and harms, or are estimates/confidence intervals quite broad?

   If you believe that your "conceptual confidence interval" around the point estimate of comparative net health benefit is limited to the boundaries of one of the four categories of comparative net health benefit above, your level of certainty is "high". "Moderate" certainty reflects conceptual confidence interval s extending across two or three categories, and may include drugs for which your conceptual confidence interval includes a small likelihood of a negative comparative net health benefit. When the evidence cannot provide enough certainty to limit your conceptual confidence interval within two to three categories of comparative net health benefit, then you have "low" certainty.

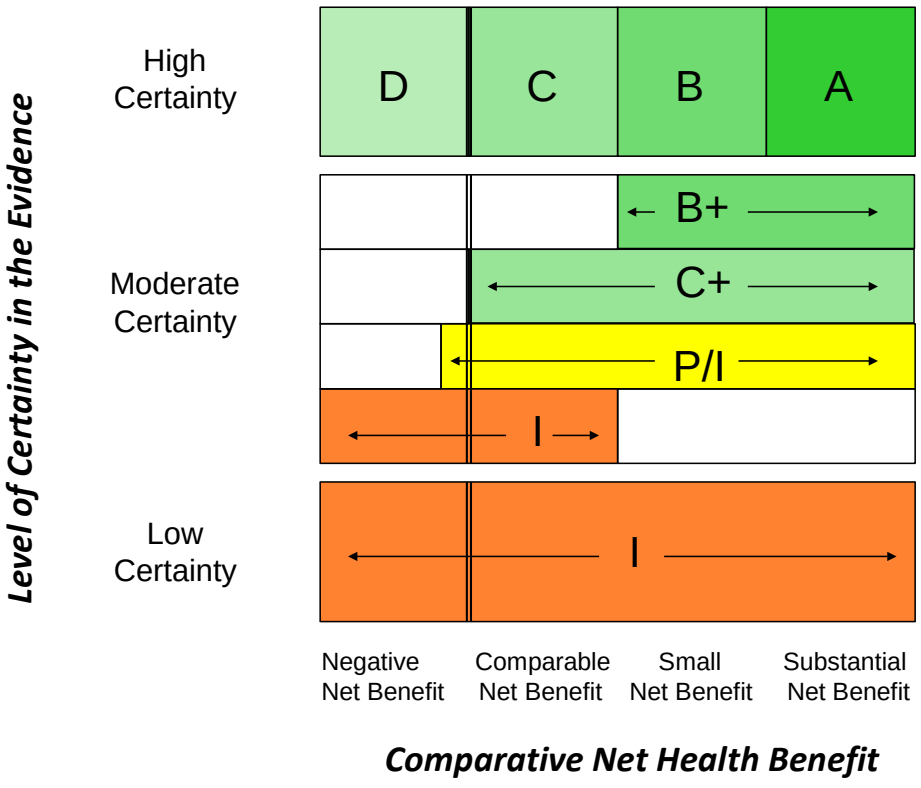
4. **Assign a joint rating in the Evidence Rating Matrix.** The final step is the assignment of the joint rating of magnitude of comparative net health benefit and level of certainty. As shown again in the figure on the following page, when your certainty is **"high,"** the estimate of net benefit is relatively assured, and so there are distinct labels available: an **A** rating indicates a high certainty of a substantial comparative net benefit. As the magnitude of comparative net health benefit decreases, the rating moves accordingly, to **B** (incremental), **C** (comparable), and finally **D**, indicating an inferior or negative comparative net health benefit for the therapeutic agent relative to the comparator.

   When the level of certainty in the point estimate is only **"moderate,"** the summary ratings differ based on the location of the point estimate and the ends of the boundaries of the conceptual confidence interval for comparative net health benefit. The ratings associated with moderate certainty include **B+** (incremental or better), which indicates a point estimate of small *or* substantial net health benefit and a conceptual confidence interval whose lower end does not extend into the comparable range. The rating **C+** (comparable or better) reflects a point estimate of either comparable, small, *or* substantial net health benefit and a lower bound of the conceptual confidence interval that does not extend into the inferior range. These ratings may be particularly useful for new drugs that have been tested using noninferiority trial designs, or those involving modifications to an existing agent to provide adherence or safety advantages.

Another summary rating reflecting moderate certainty is **P/I** (promising but inconclusive). This rating is used to describe an agent with evidence suggesting that it provides a comparable, small, or substantial net benefit over the comparator. However, in contrast to ratings **B+** and **C+, P/I** is the rating given when the conceptual confidence interval includes a small likelihood that the comparative net health benefit might actually be negative. In our experience the **P/I** rating is a common rating when assessing the evidence on novel agents that have received regulatory approval with evidence of some benefit over placebo or the standard of care, but without robust evidence regarding safety profiles when used in community practice.

The final rating category is **I** (insufficient). This is used in two situations: (a) when there is moderate certainty that the best point estimate of a drug's comparative net health benefit is comparable, but there is judged to be a moderate-high likelihood that further evidence could reveal that the comparative net health benefit is actually negative; and (b) *any* situation in which the level of certainty in the evidence is **"low,"** indicating that limitations in the body of evidence are so serious that no firm point estimate can be given and/or the conceptual confidence interval for comparative net health benefit extends across all 4 categories. This rating would be a common outcome for assessments of the comparative effectiveness of two active drugs, when there are rarely good head-to-head data available; this rating might also commonly reflect the evidence available to judge the comparative effectiveness of a drug being used for an off-label indication.

Comparative Clinical Effectiveness

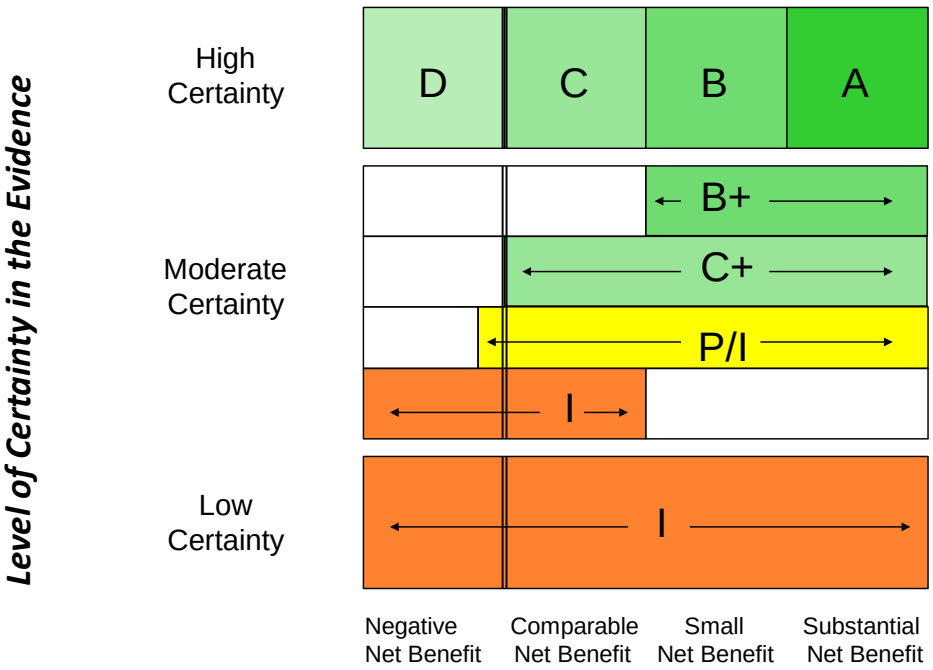


Introduction

The Institute for Clinical and Economic Review (ICER)'s Evidence Rating Matrix™ for Comparative Clinical Effectiveness is used to rate the comparative clinical effectiveness of one or more interventions relative to a comparator of interest.[1] This user's guide was developed with funding provided by the Comparative Effectiveness Research Collaborative Initiative, a joint initiative of the Academy of Managed Care Pharmacy, the International Society for Pharmacoeconomics and Outcomes Research, and the National Pharmaceutical Council (www.cercollaborative.org). Although this user's manual is focused on the application of the ICER Matrix to formulary decisions, it is designed with the flexibility to be able to compare multiple types of interventions, including drugs, devices, procedures, programs, and healthcare system processes. This method for rating comparative clinical effectiveness, depicted in the graphic below, evolved from an earlier model developed by a multi-stakeholder workgroup convened in 2007 by America's Health Insurance Plans, and relies on a joint judgment of:

a) The **magnitude** of the difference between the intervention and a comparator in "net health benefit" (the balance between benefits and risks/adverse effects); AND
b) The level of **certainty** that the body of evidence can provide in the estimates of benefits and risks/adverse effects necessary to judge the difference in net health benefit.

Comparative Clinical Effectiveness



In the sections that follow, you'll learn what all these categories and ratings mean and be able to follow a step-by-step process for assigning ratings as part of an evidence-based approach to formulary decision-making.

As a tool to support formulary decisions, the goals of the matrix are to:

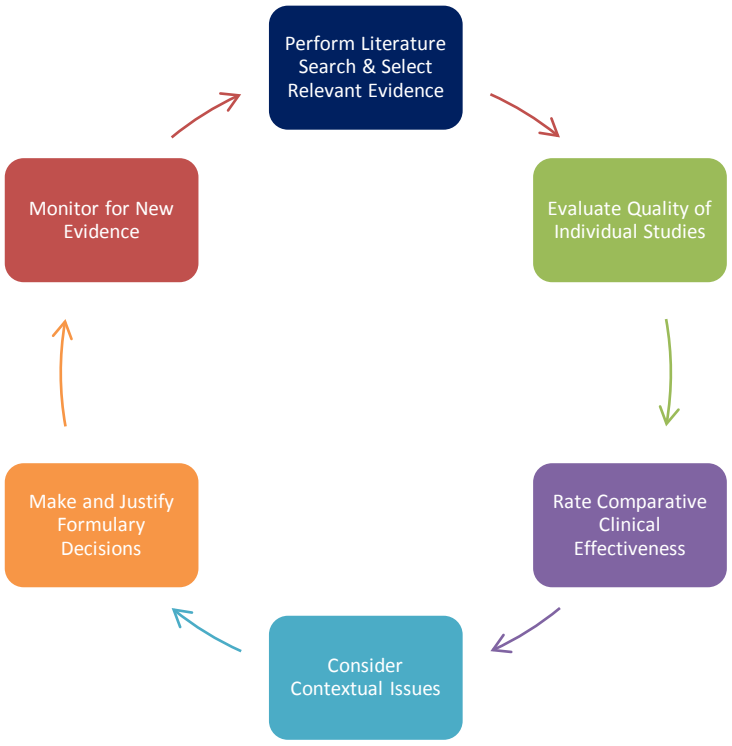
1. Provide a common approach that can be used and understood by staff of multiple disciplines and skills levels.
2. Enhance the transparency of evidence-based medical policy decisions for patients, clinicians, manufacturers, health systems, payers, and policymakers.
3. Improve the reliability of the interpretation of evidence within individual evidence review and decision-making bodies.
4. Improve the consistency of the interpretation of evidence across different evidence review and decision-making bodies.

Importantly, this matrix is **NOT** intended to be a "cookbook" algorithm leading mechanistically to a specific formulary decision. There are many additional issues that must be considered in moving from a rating of comparative clinical effectiveness to a formulary decision. For example, value judgments such as the relative importance of cost, of the severity of illness, and of clinician desire for choice must be weighed in making formulary decisions. In addition, specific benefit designs, while making use of the same evidence base, will differ based on multiple factors such as individual employer contracts, broader plan design (e.g., traditional vs. high-deductible plans), and other such concerns.

Regardless of the contextual issues, at the core of any decision should be a rigorous, transparent assessment of the evidence on comparative clinical effectiveness, and that is the function of the ICER Matrix. The process of making and then justifying these assessments, while not without challenges, helps decision-makers understand the nuances of the evidence while allowing them to move forward with using the ratings to inform their policies.

Process for Formulary Decision-Making

The ICER Matrix is designed to be one part of an overall process for evidence evaluation to support formulary decisions. This process is illustrated in the flowchart on the next page. First, to determine which studies to include within the relevant scope of the body of evidence, we recommend that the "**PICO**" framework be used; PICO is an acronym that indicates the aspects of the scope of a particular evidence review that require explicit definition, including the target **P**opulation, the **I**ntervention(s) of interest, the **C**omparator intervention(s), and the key **O**utcomes.[2] Depending on the comparison being made, it may also be relevant to address the **T**ime Horizon and **S**etting of interest (or PICOTS).



Once the PICO has been determined, and the relevant evidence obtained through literature searches, the quality of each individual study should be rated. There are many rating systems for individual study quality, and criteria vary based on the type of study. In general, however, the quality of individual studies can be assessed by considering the domains listed below, which are adapted from the methods guide of the Agency for Healthcare Research & Quality (AHRQ).[3]

Clinical Trials

- Similarity of baseline characteristics and prognostic factors between comparison groups
- Well-described methods for randomization and concealment of treatment assignment
- Use of valid, well-described primary outcomes
- Blinding of subjects, providers, and outcome assessors
- Intent-to-treat analysis (all randomized subjects included)
- Limited and non-differential loss to follow-up
- Disclosure of any conflicts of interest

Observational Studies

- Adequate sample size
- Development and use of pre-specified analysis plan
- Methods for selecting participants
- Methods for measuring risk factors or other exposures
- Methods for control of confounding and bias

- Limited and non-differential loss to follow-up
- Disclosure of any conflicts of interest

Note that while two major study types have been categorized (clinical trials and observational studies), the methods described in this guide are applicable to multiple study subtypes, including traditional and adaptive clinical trial designs, prospective cohort studies, retrospective database studies, uncontrolled case series, and model-based studies (e.g., decision analysis, simulation models).

The ICER Matrix is intended to flexibly support decision-makers in drawing conclusions from a body of evidence, regardless of each decision-maker's approach to evaluation. Some decision-makers will decide to reject studies with "significant" or "fatal" flaws that undermine their internal or external validity. For example, some Pharmacy and Therapeutics (P&T) committees will choose only to accept evidence from well-conducted randomized controlled trials (RCTs) for analyses of clinical benefits, and evidence from RCTs and observational studies of a certain minimum size or duration to evaluate safety. Others will decide to note the weaknesses and/or limitations in individual studies but keep ***all*** studies as part of the body of evidence to be considered. The ICER Matrix can be applied to all of the relevant evidence each decision-maker chooses to include in the evaluation process, including the examples and study types noted above.

After the relevant body of evidence has been selected and the quality of each individual study has been assessed, the ICER Matrix is used to assign a rating to the entire body of evidence based on the joint judgment of comparative net health benefit and the level of certainty that the body of evidence can provide in making that judgment (presented in detail in the next section). Given the possible variety of study types to be considered, the level of agreement in primary findings across study types will be a key consideration.

Once this rating for comparative clinical effectiveness has been determined, the next step in formulary decision-making is to consider contextual issues that are somewhat separate from the evidence on clinical effectiveness but certainly may affect the ultimate formulary placement. Those experienced with formulary decision-making realize that many different contextual issues complement the application of the evidence on the effectiveness and safety of a drug in formulary placement. These issues may include considerations of the cost of the drug and its comparator(s) as well as the potential cost-effectiveness of competing treatment strategies, the relative severity of the condition being treated, the opinions of key clinical and/or policy leaders, insights from the use of the drug in real-world settings, the number and acceptability of other treatment options, the possibility that a novel pharmacologic mechanism of action may confer benefit upon some patients who fail to respond to other treatments, and legal, contractual, or historical precedents.

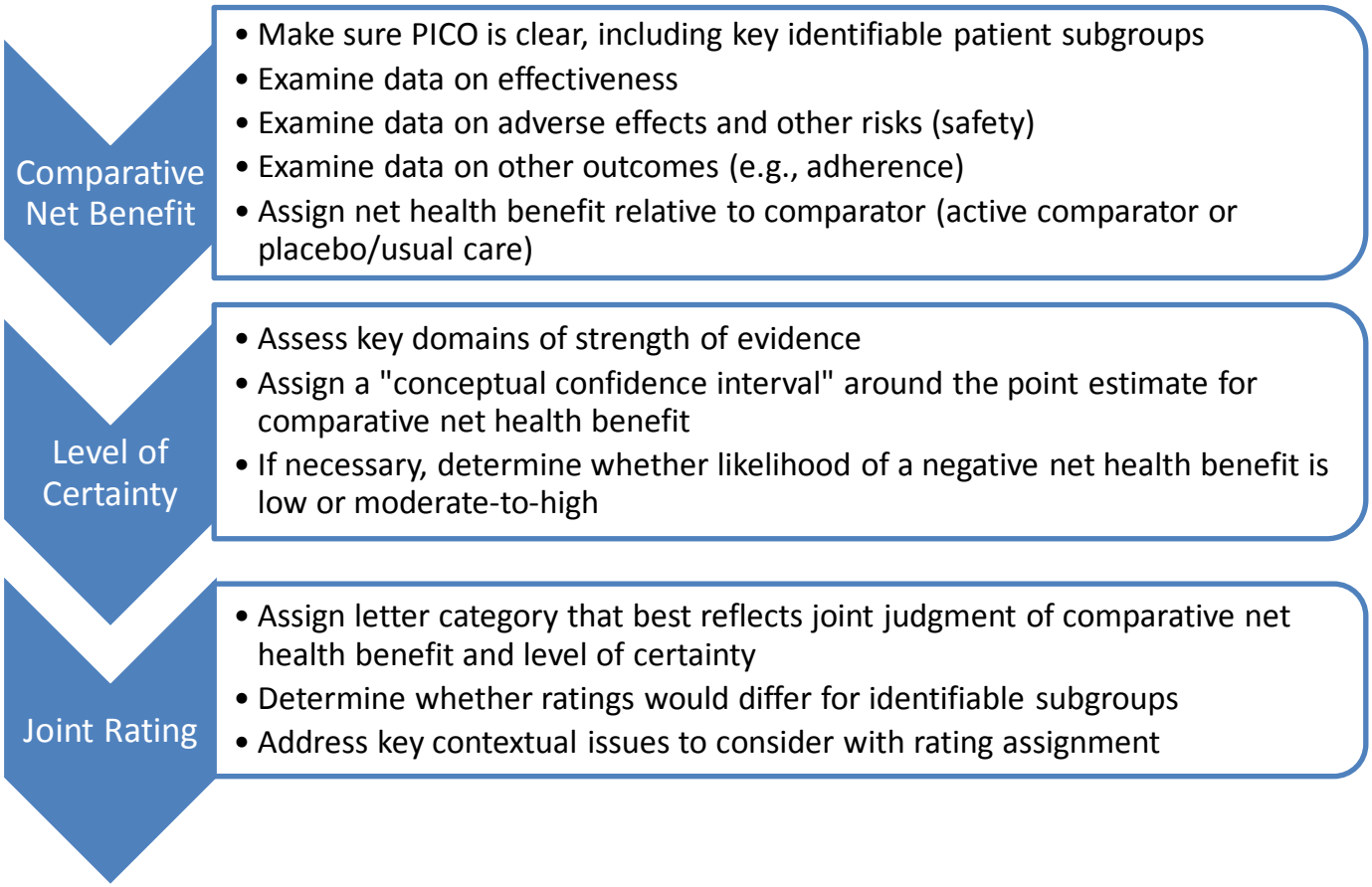
This user's guide to the ICER Matrix was developed with funding provided by the Comparative Effectiveness Research Collaborative Initiative (CER-CI), a joint initiative of the Academy of

Managed Care Pharmacy, the International Society of Pharmacoeconomics and Outcomes Research, and the National Pharmaceutical Council (http://www.npcnow.org/issue/cer-collaborative-initiative). The manual will not suggest a single pathway or "best practice" for considering and incorporating contextual considerations in a formulary decision. Some organizations may wish to create a standard protocol that will link specific contextual considerations to the ICER Matrix in arriving at a formulary decision. Other organizations will prefer to have a less structured approach. Either way, we do recommend developing an internal list of key contextual considerations and having a clearly defined process for when and how they are considered in conjunction with the ICER Matrix rating on the road to a formulary decision. This approach will improve the internal reliability and transparency of the formulary decision-making process, and may trigger important clarifications among organizational staff about the relative importance of different contextual considerations.

After making and justifying the formulary decision with reference to the ICER Matrix ratings and any contextual considerations, the final step in the overall process for producing evidence-based formulary decisions is to link the decision with a process for monitoring the literature for new evidence. Any body of evidence, and any rating of comparative clinical effectiveness, reflects only a "snapshot" in time. As new evidence emerges, it must be reviewed promptly and decisions should be made on performing a formal re-assessment of the matrix ratings.

The Rating Process

The process for assigning evidence ratings using the ICER Matrix is relatively straightforward, but requires significant thought in judging what the evidence is telling you at each step. A simple process diagram is presented below.



In the sections on the following pages, we break each of these decision steps down further so that you can understand the thought processes that go into each one.

Step 1. Magnitude of Comparative Net Benefit

Before the process of rating the body of evidence begins, it is absolutely essential to reconsider the terms of the comparison: the PICO. In most cases the PICO need not change, and the rating process can proceed. But sometimes the review of evidence has uncovered unexpected findings that suggest a change in the PICO is necessary, or that two or more separate ratings need to be developed for key identifiable patient subgroups. For example, if the original PICO suggested that the rating would compare the clinical effectiveness of two drugs for all patients with a certain condition, but the analysis of evidence has now made it clear that there are two clinically distinct subpopulations, and that the risks and benefits of the two drugs are different for these two groups, then a choice must be made whether to have two separate ICER Matrix ratings. In general, if subpopulations in which the risks and benefits differ substantially can be clinically identified at the initiation of treatment, then there should be two separate ICER ratings. In other situations there may be heterogeneity in study results that, while important, do not suggest clearly defined patient subgroups; broad regional variations in practice or differences in outcome according to certain physician characteristics, for example. In these cases, discussion of those factors driving heterogeneity should accompany the overall ICER rating.

Two additional considerations for re-affirming the PICO are important to keep in mind. The first is whether the evidence is more plentiful and/or more persuasive for some patient groups than for others. Evidence might be robust for patients under age 65, for example, but sparse or non-existent for older patients. It may be advisable to consider performing separate ratings for these different patient groups. The second consideration involves the comparisons made in the evidence base. It might be the case that there is evidence for a new drug comparing it to one or more active comparators already on your formulary, in addition to the expected evidence comparing the new drug to placebo. Again, it may be advisable to apply the rating system separately to each type of evidence. By doing so, you will be able to inform (a) the new drug's general formulary placement; and (b) its relative formulary standing vs. existing alternatives.

Once the PICO is re-established, the first step in arriving at an overall rating is to judge the comparative net health benefit of the drug of interest relative to its selected comparator. In other words, considering the evidence on safety and clinical benefits for the drug of interest and for the comparator, how does the net health benefit for the drug of interest compare overall to that of the comparator?

When making this judgment of the comparative net health benefit of the drug and the comparator, there are four possible categories within the ICER Matrix. These categories are listed below:

- ***Negative:*** the drug produces a net health benefit inferior to that of the comparator
- ***Comparable:*** the drug produces a net health benefit comparable to that of the comparator
- ***Small/Incremental:*** the drug produces a small positive net health benefit relative to the comparator
- ***Substantial:*** the drug produces a moderate-to-large positive net health benefit relative to the comparator

You can think of picking one of these categories for the magnitude of comparative net health benefit as your best possible "point estimate" given the existing evidence. Don't worry yet about how strong or weak the evidence seems on effectiveness or safety – that will come into play in the next stage of the rating process. At this step, for example, it may be the case that you will rate the comparative net health benefit as "comparable" based on evidence from high-quality head-to-head RCTs, but a "comparable" rating might also be your best guess at the point estimate based on indirect evidence from a handful of drug vs. placebo studies combined through a meta-analysis.

It is worth noting that there are two variants of these categories available to users: "comparable or better" and "incremental or better". While these categories also must take into consideration the level of certainty in the point estimate (see the next section), they are also of great utility when considering a newly-approved medication. For example, some new medications may be structurally identical to existing alternatives but with simpler dosing schedules or more convenient drug delivery, suggesting clinical performance that is at least as good as, and perhaps incrementally better than, existing treatments. In other situations, a new drug may offer a distinct advantage over existing treatments, but the true level of incremental benefit (i.e., small vs. substantial) is not yet known.

Conceptually, "net" health benefit requires a judgment of the balance of all benefits, adverse effects, and other risks. To assess this balance the evidence on clinical benefits and safety should generally be evaluated separately to ensure that evidence on safety receives an appropriate focus. Evidence on clinical benefits related to the outcomes specified in the PICO should be assessed, with equal attention paid to potential weaknesses in the body of evidence. Then, evidence should be reviewed and summarized regarding important safety issues for both the drug of interest and the comparator, including common side effects and rare adverse events. Limitations to the strength of evidence on safety should be noted. We will provide further guidance on methods to

assess and summarize the strength of the body of evidence on clinical benefits and safety in the "Level of Certainty" section of this guide.

Once the evidence on clinical benefits and safety for the drug of interest and the comparator has been assessed, a comparison of their relative "net" health benefits should be made. To accomplish this comparison in a robust, transparent fashion, we recommend that the magnitude of benefits and the magnitude of adverse effects and other risks be analyzed and then summed up quantitatively across all studies in the body of evidence whenever possible. For example, comparisons could be made between summary estimates for each drug of the percentage of patients experiencing a certain level of side effects. Wherever possible, these comparisons should focus on ***absolute*** rather than relative risks, so that the potential for the drug of interest and its comparator to cause harm is transparent.

It is also often the case that other outcomes that are not neatly categorized as "benefits" or "risks" should be considered in forming an overall impression of net health benefit. For example, a drug-drug interaction might modulate the effectiveness of a given therapeutic agent, cause a serious adverse effect, or both. Patient adherence to medication also might be of concern, due to differences in drug formulation, route of administration, dosing, or other issues. Consideration of additional outcomes such as these can help round out the comparisons between drugs, and in some cases might tip the balance between them.

But benefits, risks, and other outcomes must be weighed together in some fashion in order to arrive at judgment of comparative net health benefit, and there are different ways this can be done. When a drug and its comparator have the exact same types of clinical benefits and risks, and the only difference is the rate of these outcomes, then a quantitative comparison is relatively straightforward. However, competing drugs often have not only differing levels of the same benefits and risks but different *types* of benefits and risks. For example, one effective drug for rheumatoid arthritis might cause fatigue and present a risk of serious infection, whereas its comparator is somewhat less effective at improving joint symptoms but causes less fatigue; the drug does not pose an infection risk but does carry a small risk of serious liver disease. How can the comparative "net" health benefit of these two drugs be compared?

Some organizations will choose to construct a formal mathematical comparison of benefits, risks, and/or other outcomes. Examples include measures of absolute or relative reduction in risk as well as companion measures such as number needed to treat (NNT) to achieve an additional treatment "success" or number needed to harm (NNH), a measure of the number needed to treat before an additional serious adverse event is experienced.[4,5] Still other analyses explicitly combine data on benefits and risks into summary measures. Examples include health utility analysis, which produces an overall estimate that combines benefit and risk known as the quality-adjusted life

year (QALY); incremental net health benefit (INHB) analysis, which separately summarizes benefits and risks and measures the difference between the two; and multi-criteria decision analysis (MCDA), which allows for measurement of benefit and harm along a continuum of attributes that can be weighted to reflect their relative importance.[6-8] The magnitude of the comparative net health benefit can then be judged by comparing these measures between drug strategies. The more effective arthritis drug with a small risk of serious infection might, for instance, be estimated to produce an average QALY gain of 0.1 in relation to its comparator, suggesting it produces a greater net health benefit. Even with this approach, however, judgment is required regarding the clinical significance of varying ranges of QALY differences. Is the QALY difference of 0.1 clinically insignificant, or is it a "small" or even a "substantial" difference in net health benefit? Internal discussions and external benchmarking are the only way to develop a transparent, reliable approach to making this decision.

Whereas some P&T committees will prefer to use a common metric such as those described above, in our experience most groups prefer to judge the comparative overall balance of benefits and risks implicitly. If you take this approach, benefits and risks are not combined quantitatively, and their relative importance remains a value judgment. The arthritis drug that is safer but somewhat less effective at reducing joint symptoms might thus be judged to offer a "comparable" net health benefit, or it might be judged to offer a "small" positive net health benefit. Indeed, it could even be judged to offer an "inferior" net health benefit. As you can tell, whether benefits and risks are summed and compared quantitatively or qualitatively, value judgments will still be a part of the final ICER rating. There is no quantitative definition of the boundaries between the four categories used in the ICER Matrix to rate the comparative net health benefit. Efforts to quantify the relative benefits and risks with as much precision as possible will help bring these boundaries into sharper relief, but individual formulary decision-makers will have to decide for themselves where to draw the line. The ICER Matrix is a tool to clarify the inputs to the decision, and to structure the debate and discussion within a P&T committee so that these judgments and the resulting final ratings can be made as transparently and reliably as possible across all decisions.

Although judgment remains an important component of the rating system, this does not imply that the rating system is entirely subjective and that there can be no way to calibrate decisions within and across P&T committees. It may be useful to consider some prototypical examples featuring actual drug comparisons that would lead to each category of net health benefit (as well as the "variants" we discussed earlier). Consider the following examples:

- **Negative**
  - *Aspirin vs. Warfarin for Stroke Prevention in Patients with Atrial Fibrillation and a Moderate-High Risk of Stroke.* Evidence from multiple clinical trials demonstrates

that aspirin poses similar risks of bleeding compared to warfarin but is only approximately half as effective in preventing strokes.

- **Comparable**
  - *ACE Inhibitors (ACEIs) vs. Angiotensin Receptor Blockers (ARBs) for Long-term Control of Hypertension.* Evidence from multiple clinical studies and systematic reviews indicates no differences in effectiveness between ACEIs and ARBs, and only small differences in minor side effects.
  - **Comparable or Better→***Controlled-release vs. Immediate-Release Formulations of Methylphenidate in ADHD.* Data from multiple randomized comparisons of immediate-release methylphenidate (e.g., Ritalin®) vs. controlled-release forms (e.g., Concerta®, Metadate®) suggest no differences in symptom control or adverse effects between formulations, but longer-term observational studies suggest adherence advantages for controlled-release medications.
- **Small/Incremental**
  - *Tissue Plasminogen Activator (t-PA) vs. Streptokinase for Myocardial Infarction.* Evidence from multiple clinical trials suggests a small but statistically-significant difference in all-cause mortality in favor of t-PA, balanced against an increased risk of hemorrhagic stroke with t-PA.
  - **Incremental or Better→***S-Ketamine vs. Racemic Ketamine for Operative Anesthesia.* In contrast to data suggesting that isomeric compounds are generally no more effective than their racemic counterparts, S-ketamine (Ketanest® S) has been found to produce similar levels of anesthesia and analgesia in comparison to racemic ketamine (e.g., Ketanest®, Ketalar®); however, S-ketamine is also eliminated more quickly by the body, and has been associated with lower rates of agitation, hallucination, and other anesthesia reactions.
- **Substantial**
  - *Imatinib (Gleevec®) vs. Interferon in Chronic Myelogenous Leukemia.* Evidence from pivotal clinical trials indicates a rate of complete treatment response for Gleevec 20 times greater than that for interferon, as well as a much lower rate of treatment withdrawal due to adverse events.

Over time, repeated use of this rating approach will create an internal track record of judgments derived from the available evidence. This is turn may be used as a kind of "common law" to guide future judgments, which should ultimately lead to improved internal reliability in translating the major study findings into judgments regarding net health benefit.

Step 2. Level of Certainty

After the categorical "point estimate" (i.e., substantial, small, comparable, or negative) of the comparative net health benefit has been determined, the next step involves making a judgment about the level of certainty in that point estimate provided by the available evidence. One way to view this step is that it involves deciding what your "conceptual confidence interval" is around your earlier point estimate of the magnitude of the comparative het health benefit.

A broad conceptual confidence interval would imply a lower level of certainty in your point estimate of the comparative net health benefit; for example, you might consider drug A to be comparable to drug B, but issues with the evidence suggest that you cannot rule out that drug A might provide a small benefit or even be inferior to drug B. Less certainty could arise because you have questions regarding the evidence on comparative clinical benefits, safety, or both. So how should you evaluate the entire body of evidence to make this judgment about the level of certainty?

ICER follows methods developed by the United States Preventive Services Task Force (USPSTF), AHRQ, and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group.[3,9-10] Links to each organization's full methods documentation are available in the Appendix to this guide. While each organization has some differences in its approach to rating the strength of evidence, there is a fundamental similarity: the body of evidence should be evaluated, in its entirety, along a series of criteria or "domains." Specific labels of these domains differ across USPSTF, AHRQ, and GRADE; however, the most important domains are considered in all of these systems, as presented below and on the next page.

- Bias:
  - The risk of bias from the study designs within the body of evidence (e.g., "channeling bias" in comparative observational studies of two drugs)
- Applicability:
  - The generalizability of the patients and clinicians to "real-world" populations (e.g., differences in patient populations between traditional and pragmatic clinical trial designs, inclusion of real-world database studies in body of evidence)
- Consistency:
  - The degree to which different studies have similar key findings

- Directness:
    - The relative directness vs. indirectness of the comparison of drug vs. comparator possible from the available evidence; and/or
    - The relative directness with which measured outcomes in the studies, e.g. surrogate outcomes, reflect true patient-centered outcomes
- Precision:
    - How precise the results on key outcomes are in available evidence

Essentially, the greater the level of comfort in these domains (i.e., low risk of bias and high confidence in the other domains), the greater the strength of evidence, and the higher the level of certainty in what the evidence is telling you. Other domains may also be relevant in evaluating level of certainty depending on the drugs evaluated. These other domains include:

- Evidence of a dose-response relationship
- Biologic plausibility of the treatment effect
- Publication bias (e.g., are the results primarily from small, single-center studies vs. large, multicenter evaluations?)

We do not use any algorithm based on these domains to judge the level of certainty as part of the ICER Matrix rating. Comments in an evidence review should formally discuss how the body of evidence stacks up in each domain. Some groups choose to use a summary score for each domain; however, even if this approach is taken we do not recommend trying to devise a mathematical equation using domain scores to arrive at a final summary score for level of certainty. Sometimes the issue of precision will have much greater weight in a final judgment of level of certainty than other domains; in other cases it may be the consistency of findings that matters most. For example, even if the entire body of evidence is comprised of 50 case series, if every study, among a broad variety of institutions and patient populations, generates very similar results, then the consistency of the body of evidence may give you a high level of certainty in your point estimate of comparative net benefit, even if the "quality" of each of the individual studies is poor due to risk of bias.

A mathematical equation using scores from all domains cannot capture these kinds of situations, and could, in an attempt to bring greater reliability to the process, actually yield invalid judgments. As with the ultimate judgment on comparative net health benefit, equations and algorithms can help to a certain extent. For example, findings from a simulation model might help clarify the key drivers of a difference in a summary measure of effectiveness – specific types of side effects, issues with medication compliance, etc. Ultimately, however, a transparent justification of the final judgment is superior to a blind application of a quantitative approach; as noted previously, a small difference in a quantitative score may be considered clinically significant in multiple

directions, or not significant at all, so documentation is necessary to illuminate the rationale for a given rating. Disagreements about the assessment of each domain and of the overall level of certainty are certainly likely, even among reviewers in the same group. But in such cases this framework of domains can help clarify the source of the different opinions regarding the body of evidence and lead to more focused discussions seeking consensus.

The ICER Matrix has three levels of certainty: "High", "Moderate", and "Low". As with judgments regarding the boundaries of the categories for comparative net health benefit, there are no exact quantitative methods for defining the boundaries between these categories. However, there are typical characteristics of bodies of evidence that are representative of each level of certainty, as illustrated in Box 1 on the next page.

"Conceptual Confidence Intervals"

There is another way to think about how to judge the level of certainty. We find it useful to consider that conceptual confidence intervals around a point estimate that do not extend beyond a single box of comparative net health benefit represent a "high" level of certainty. Thus, if your point estimate of comparative net health benefit is "comparable," and you feel that the reasonable bounds of your conceptual confidence interval do not extend into either "negative" or "small," then you have high certainty in a "comparable" net health benefit.

"Moderate" certainty reflects conceptual confidence interval s extending across two or three categories, and may include drugs for which your conceptual confidence interval includes a small likelihood of a negative comparative net health benefit. For example, this may arise with newly-approved drugs with evidence of clinical benefit but unanswered safety questions that require further study post-marketing. In these situations, you must ask yourself: "Do I believe that the likelihood that further evidence will move the net health benefit to negative is small, or is it moderate-to-large?" If you consider this possibility to be small, your certainty in the net health benefit is "moderate." Moderate certainty also applies in situations where the evidence suggests that the true net health benefit lies between comparable and negative only. Such situations might arise for a drug with comparable outcomes based on noninferiority trials but with open questions regarding its long-term net health benefit relative to existing therapies.

When the evidence cannot provide enough certainty to limit your conceptual confidence interval within two to three categories of comparative net health benefit, then you have "low" certainty. This implies that there really is no way to estimate the level of net health benefit within any sort of conscripted range due to limitations in the amount and/or quality of available evidence.

Box 1. General Characteristics of Evidence Providing Different Levels of Certainty.

High

- Mostly high-quality, larger studies
- Conducted in representative patient populations
- Direct comparisons available
- Address important outcomes or validated surrogate outcomes
- Long-term data on benefits/risks available
- Consistent results
- Future studies unlikely to change conclusions

Moderate

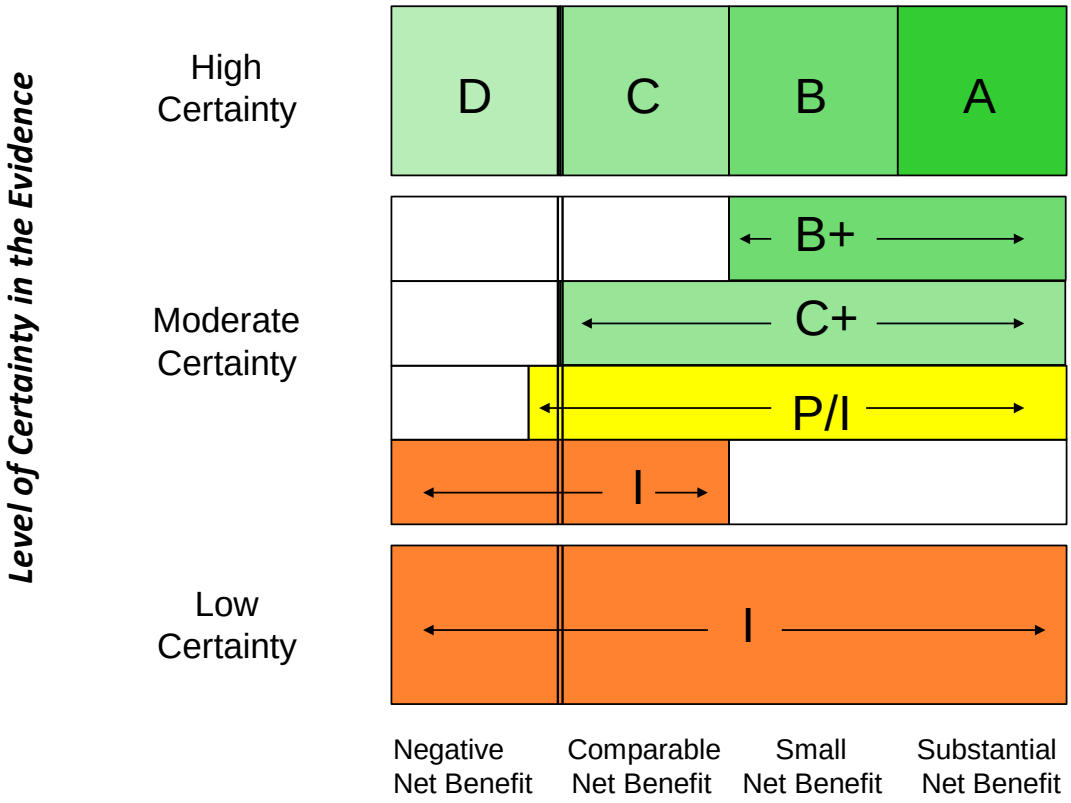
- Mix of study quality
- Cannot estimate net benefit with good precision, based on limitations including:
    - Weak study design or conduct
    - Inconsistent findings
    - Indirect evidence only
    - Limited applicability of results
    - Evidence of reporting bias
- Future studies may result in modest shifts in estimates of net health benefit

Low

- Mostly poor-quality, smaller studies
- Evidence insufficient to estimate net benefit at all
- Flaws in evidence base make it impossible to determine if intervention inferior, comparable, or superior to comparator
- High likelihood that new evidence would substantially change conclusions regarding net benefit

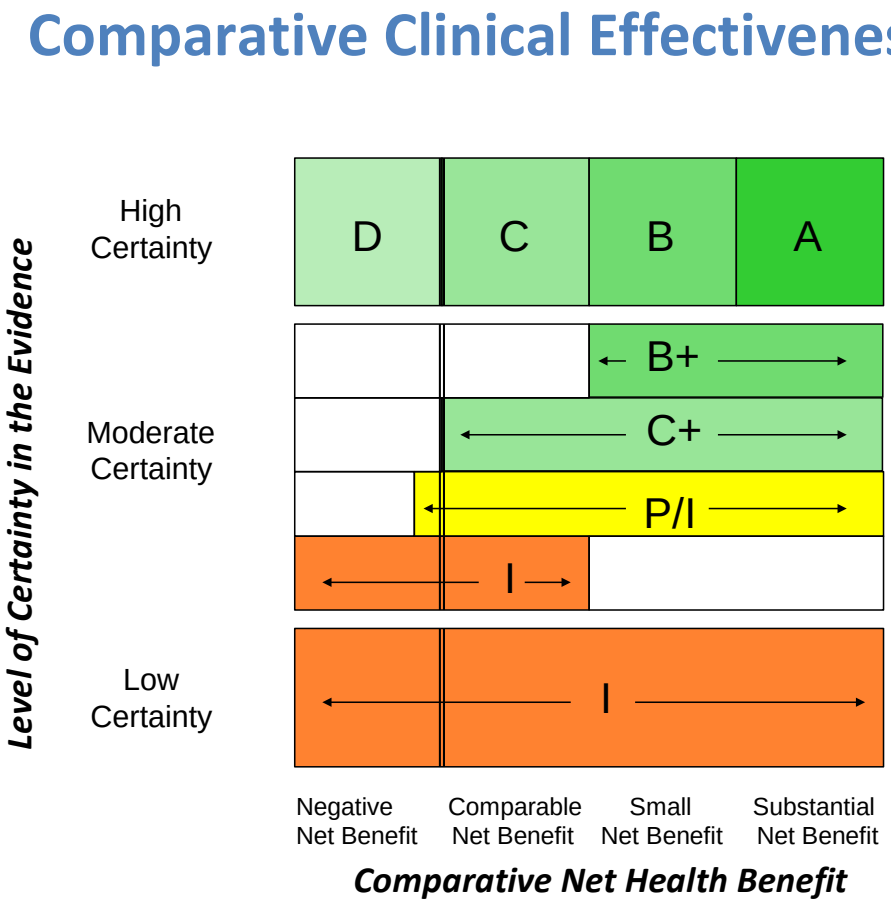
To get a better feel for this, try a thought experiment: take a look again at the ICER Matrix below and pick a few potential point estimates for comparative net health benefit. Then, for each one try out different spans of conceptual confidence intervals, some that stay in one box, others that extend for one or more on either side. You'll soon get the general idea for how the joint consideration of comparative net health benefit and level of certainty work together to lead to a combined ICER rating. And we are now ready to examine this final step in the process.

Comparative Clinical Effectiveness



Step 3. Joint Rating

It is time to turn our attention back to the ICER Matrix to discuss the summary ratings used to describe the comparative clinical effectiveness of a therapeutic agent versus its comparator. The matrix is illustrated again below.



A = "Superior" - High certainty of a substantial (moderate-large) net health benefit
B = "Incremental" - High certainty of a small net health benefit
C = "Comparable"- High certainty of a comparable net health benefit
D="Negative"- High certainty of an inferior net health benefit

B+="Incremental or Better" – Moderate certainty of a small or substantial net health benefit, with high certainty of at least a small net health benefit
C+="Comparable or Better"-Moderate certainty of a comparable, small, or substantial net health benefit, with high certainty of at least a comparable net health benefit
P/I = "Promising but Inconclusive " - Moderate certainty of a comparable, small, or substantial net health benefit, and a small (but nonzero) likelihood of a negative net health benefit

I = "Insufficient" – Either moderate certainty that the best point estimate of comparative net health benefit is comparable or inferior; or any situation in which the level of certainty in the evidence is low

Not surprisingly, the greatest level of precision in these ratings comes when there is a high level of certainty. In this situation, your best point estimate of net benefit is relatively assured, and so there are distinct labels for each level of benefit. An "A" rating (superior) indicates a high certainty of a moderate-to-large comparative net health benefit relative to the comparator. As the magnitude of comparative net health benefit decreases, the rating moves accordingly, to "B" (incremental), and "C" (comparable). Finally, the "D" rating indicates a clearly inferior or negative comparative net health benefit for the drug relative to the comparator.

When the level of certainty in the point estimate is only moderate, however, the summary ratings change to reflect that the conceptual confidence intervals around point estimates of comparable, small, or substantial net health benefit are not as precise. As discussed previously, if available evidence suggests the drug of interest provides at least a small benefit and you believe it may in fact provide a substantial benefit, you would use a "B+" rating (incremental or better); similarly, if evidence suggests that a new agent is at least comparable to an alternative therapy but there are reasons to believe it may well be better, with very good reasons to believe that it is not inferior, then the rating would be "C+" (comparable or better). The "P/I" (promising but inconclusive) category describes a therapeutic agent with evidence suggesting that it provides a comparable, small, or substantial net benefit over the comparator. This point estimate, however, includes a small but not unreasonable chance that the true comparative net health benefit is "inferior." The P/I rating is a particularly important one for formulary decision-making, as evidence for new drugs evaluated by regulators under so-called "expedited review" conditions suggests benefits over placebo and potentially over active comparators as well, but limitations in the evidence, particularly on longer-term safety and effectiveness, tend to reduce certainty in the true magnitude of comparative net health benefit.

The final rating category is "I" (insufficient). This is used in two situations: (a) when there is only moderate certainty that the best point estimate of a drug's comparative net health benefit is comparable or inferior only; and (b) any situation in which the level of certainty in the evidence is low, indicating that it is problematic to give a specific point estimate for comparative net health benefit and that limitations in the body of evidence are so severe that the conceptual confidence interval extends across multiple categories in the ICER Matrix, including negative net health benefit.

In assigning these ratings, the important thing to remember is that any summary rating should be coupled with a stated rationale and justification to make the process as transparent as possible to all stakeholders. Nevertheless, the joint rating of magnitude of net health benefit, likelihood of a negative net health benefit, and level of certainty in the evidence can have great value for P&T committees. For one, the rating matrix offers a common approach that, over time, can be reliably applied by frontline reviewers within an organization. In addition, organizations that are so

inclined can map these ratings directly to a suite of possible formulary decisions. Finally, the summary rating categories are clear and well-understood, which enhances the likelihood of consistent decision-making across organizations.

We hope that you find the rating system useful and of benefit to your organization. On the following pages you will find case studies documenting real-life as well as hypothetical applications of the rating matrix to put the use of the matrix in further context. More information on ICER and the ICER Matrix can be found at www.icer-review.org.

The ICER Matrix in Action

To illustrate how the ICER Evidence Rating Matrix can be used to support formulary decisions, it is useful to review case studies from prior ICER work: judgments regarding net benefit and level of certainty, the summary rating applied, and the rationale given for the rating. The two case studies below come from an ICER appraisal of management options for atrial fibrillation (AF).[11]

Case Study #1: Dabigatran for Stroke Prevention in Atrial Fibrillation

The Evidence: Dabigatran (Pradaxa®, Boehringer Ingelheim Pharmaceuticals, Inc.) is a direct thrombin inhibitor that is FDA-approved for use as an anticoagulant in patients with AF not caused by valvular disorders. It is an alternative to warfarin, which has been considered the standard of care for stroke prevention in AF for over two decades. Dabigatran's approval was based on the conduct of a single RCT, the RE-LY study,[12] which randomized over 18,000 patients to receive one of two doses of dabigatran or warfarin and followed them for two years. Both doses of dabigatran were associated with statistically-significantly lower rates of hemorrhagic stroke vs. warfarin, and the higher dose was also associated with reduced rates of total stroke and vascular mortality. Rates of major bleeding, the primary side effect of all anticoagulants, was comparable or lower with dabigatran relative to warfarin. However, dabigatran was associated with a statistically-significantly higher rate of myocardial infarction (MI) vs. warfarin, a finding that was unexplained at the time of approval and remains under investigation today.[13]

ICER's Rating: Although a single RCT would not usually provide enough certainty to merit anything other than an "Insufficient" rating, ICER's judgment was that RE-LY produced unusually persuasive findings. The study was very large, well-designed, and produced highly consistent findings across drug doses. We judged the level of certainty to be moderate in a small to substantial net health benefit for dabigatran, the "promising but inconclusive" (P/I) rating, based on the findings of this trial as well as other advantages vs. warfarin (e.g., fixed dosing, no monitoring requirements). We did not raise the level of certainty to "high", however, as questions remained about the possible higher comparative risk of MI. However, given known safety concerns with warfarin as well as the clinical advantages apparent for dabigatran, it was felt that the likelihood of further evidence on dabigatran safety translating into a negative net health benefit was relatively low, leading to the P/I rating.

Since dabigatran's approval, meta-analyses have continued to raise concerns about dabigatran's cardiovascular safety;[13,14] in addition, real-world adverse event data suggest that serious bleeding with dabigatran and other direct thrombin inhibitors might be problematic, if not impossible, to reverse in some patients.[15] Certainty in net health benefit is likely to remain "moderate" until further evidence addresses these questions.

Case Study #2: Dronedarone for Rhythm Control in Atrial Fibrillation

The Evidence: Dronedarone (Multaq®, Sanofi-Aventis L.L.C.) is an anti-arrhythmic agent that is FDA-approved to reduce the risk of AF-related hospitalization in patients with the paroxysmal or persistent forms of the condition who do not have recently-decompensated heart failure and/or permanent AF. Dronedarone has been tested in multiple large multicenter placebo-controlled RCTs and one head-to-head comparison (the DIONYSOS study) with amiodarone, which is widely believed to be the most effective agent at restoring normal sinus rhythm.[16] In this trial, recurrence of AF occurred at a rate more than 20% higher with dronedarone (63.5%) than with amiodarone (42.0%). This finding was consistent with the results of a mixed treatment comparison performed in the ICER appraisal, in which dronedarone was found to be 70% less likely to restore normal sinus rhythm than amiodarone, and was no more effective than other anti-arrhythmic agents such as sotalol. However, use of amiodarone has been associated with relatively high rates of thyroid and pulmonary toxicity, which can be permanent in some cases; the occurrence of these events with dronedarone has been very low.

ICER's Rating: Although results are available from only one major head-to-head RCT of dronedarone and amiodarone in AF, ICER felt that the findings of this study, when combined with evidence accumulated from RCTs of these drugs vs. placebo, allows a high level of certainty that the comparative net health benefit of the two agents is essentially "comparable" (C). This rating was due to the central tradeoff apparent from the evidence: dronedarone is far less effective than amiodarone at maintaining patients in normal sinus rhythm, but it also offers a lower risk of serious long-term toxicity. Real-world data available after the ICER review suggest that dronedarone's performance at maintaining normal sinus rhythm is suboptimal compared to even other first-line agents, however, and that dronedarone is associated with adverse effects serious enough to warrant discontinuation in approximately one-quarter of those who receive it.[17] In addition, new questions have been raised regarding the drug's association with cardiovascular morbidity and mortality.[18,19] If additional evidence is consistent with these findings, the balance of benefits and risks might shift enough to warrant changing dronedarone's rating to "insufficient" (I) (i.e., the true point estimate of benefit lies between inferior and comparable) or even "inferior" (D) in comparison to amiodarone.

Case Study #3: Leukotriene Inhibitors in Mild-Moderate Asthma (Hypothetical)

In addition to case studies based on actual ICER experiences, it may also be useful to consider how the ICER rating matrix might be applied in a different clinical area. Take a look at the description below and see what your decision would be.

The Evidence: Multiple clinical trials have demonstrated that inhaled corticosteroids provide control of asthma symptoms superior to that of leukotriene inhibitors in children and adults with mild-to-moderate chronic asthma. A recent systematic review of 27 RCTs indicated that patients receiving leukotriene inhibitors were significantly more likely to experience exacerbation of asthma symptoms;[20] the NNH for leukotriene inhibitors was 26 (i.e., 26 patients receiving leukotriene inhibitors instead of inhaled corticosteroids would cause one additional exacerbation). Inhaled corticosteroids also showed significant improvements in forced expiratory volume, symptom-free days, use of rescue medications, nocturnal awakenings, and quality of life. There were no major differences in drug safety.

This evidence is balanced, however, against data from real-world observational studies suggesting that adherence to leukotriene inhibitors, which come in tablet form and have simple dosing schedules, is far better than that of inhaled corticosteroids, which involve complex delivery systems that may be difficult to manage for children, and are often dosed multiple times daily. A recent study of nearly 30,000 children using administrative databases in Quebec found that the number of days with "controller medication in hand" among leukotriene inhibitor recipients was twice that of patients receiving inhaled corticosteroids.[21] There was no difference in the rate of exacerbations between groups among children with an exacerbation history, and an exacerbation rate among inhaled corticosteroid users more than *twice* that of leukotriene inhibitor recipients without an exacerbation history.

The Rating: Considering only data from RCTs, the choice appears to be clear. Inhaled corticosteroids provide significant clinical benefits relative to leukotriene inhibitors, without any major safety concerns. "Superior", right? But compliance is relatively poor with inhaled corticosteroids, so how likely are these benefits to actually be realized? What would you do? If you only consider RCT data in making your decision, you might retain "Superior" as your rating. If you find the real-world findings compelling, however, you might downgrade to "Small" or maybe even "Comparable", given what appears to be a substantial adherence advantage for leukotriene inhibitors. You might also consider the "B+" or "C+" ratings for situations in which adherence support might be provided for inhaled corticosteroids.

Whatever you decide, remember to fully document your rationale and justification!

Case Study #4: Fidaxomicin vs. Vancomycin for C. difficile Infection (Hypothetical)

The Evidence: Infection with the C. difficile bacterium is common in hospitalized and/or immunocompromised patients, particularly after treatment with broad-spectrum antibiotics for other infections. "C-diff" infection causes diarrhea and may also result in dangerous inflammation in the colon, and typically requires long-term antibiotic treatment in all but the mildest of cases. Moderate-to-severe cases, as well as patients who have relapsed after an initial course of antibiotic therapy, are typically treated with vancomycin (Vancocin®, ViroPharma Inc.). Vancomycin therapy, while more effective than initial treatment, nevertheless carries a relatively high rate of relapse (up to 25%) as well as a risk of antibiotic-resistant disease.

Fidaxomicin (Dificid™, Optimer Pharmaceuticals, Inc.) is a new macrolide antibiotic recently approved to treat C. difficile infection. In two Phase III noninferiority trials comparing 10 days of oral therapy with fidaxomicin and vancomycin, cure rates approached 90% for both drugs and did not differ statistically.[22,23] Rates of treatment-related adverse events were also similar between groups. In one of these trials, the rate of recurrence at 4 weeks was also assessed and found to be statistically-significantly lower with fidaxomicin (15.4% vs. 25.3% for vancomycin, p=.005).[22] Fidaxomicin is also simpler to dose, as it is typically given twice a day vs. 4 times daily for vancomycin. The drug has not yet been studied in patients with refractory or relapsing disease, however, and the recurrence advantage observed was primarily in patient with less virulent strains of C. difficile. In addition, safety and toxicity has been evaluated in fewer than 1,000 patients at this point.

The Rating: The noninferiority nature of these pivotal trials lends itself to consideration of a "comparable or better" rating (C+). The evidence accumulated to date suggests that, in patients with new-onset illness, rates of both initial cure and adverse events are comparable for fidaxomicin and vancomycin. There are even data to suggest that fidaxomicin might result in lower rates of recurrence than vancomycin, and dosing is simpler for the patient, so there is a real possibility of a positive net health benefit. Some users might in fact consider the recurrence data compelling enough to rate fidaxomicin as "incremental or better" (B+). On the other hand, safety data are limited and fidaxomicin has not yet been adequately studied in the same difficult-to-treat patients typically reserved for vancomycin in clinical practice. For some users, this might translate into a small possibility that fidaxomicin is inferior to vancomycin. You guessed it – P/I, at least until additional evidence is available.

Again, fully documenting your rationale and justification for your ratings will allow others to understand the thought behind the ratings!

References

1. Ollendorf DA, Pearson SD. An integrated evidence rating to frame comparative effectiveness assessments for decision makers. *Med Care* 2010;48:S145-S152.

2. Haynes R: Forming research questions. In Clinical Epidemiology: How to do Clinical Practice Research. 3rd edition. Edited by Haynes R, Sacket D, Guyatt G, Tugwell P. Philadelphia, PA: Lippincott Williams & Wilkins; 2006:3-14.

3. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(12)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. April 2012.

4. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* 1997;126:712–20.

5. Sedgwick P. Statistical question: number needed to harm. *BMJ* 2011;342:d2811.

6. Brazier JE, Ratcliffe J, Tsuchiya A, Salomon J. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press, 2007.

7. Garrison LP Jr, Towse A, Bresnahan BW. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. *Health Affairs* 2007;26:684–95.

8. Thokala P. Multiple criteria decision analysis for health technology assessment: report by the decision support unit. Sheffield, UK: School of Health and Related Research, University of Sheffield. February, 2011.

9. Sawaya GF, Guirguis-Blake J, LeFebre M, et al. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med* 2007;147:871–75.

10. Guyatt GH, Oxman AD, Vist G, et al. for the GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-26.

11. Ollendorf DA, Silverstein MD, Bobo T, Pearson SD. Final Appraisal Document: Rhythm Control and Stroke Prevention Strategies for Patients with Atrial Fibrillation. Boston, MA: Institute for Clinical and Economic Review, September 2010. Available at: http://www.icer-review.org/index.php/Completed-Appraisals/a-fib-appraisal-1209.html. Accessed October 29, 2012.

12. Connolly SJ, Ezekowitz MD, Yusuf S, et al. for the RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2009;361:1139-51.

13. Uchino K, Hernandez AV. Dabigatran association with higher risk of acute coronary events: meta-analysis of noninferiority randomized controlled trials. *Arch Intern Med* 2012;172:397-402.

14. Mak KH. Coronary and mortality risk of novel oral antithrombotic agents: a meta-analysis of large randomised trials. *BMJ Open* 2012;2:e001592 doi:10.1136/bmjopen-2012-001592.

15. U.S. Food and Drug Administration. Pradaxa (dabigatran etexilate mesylate): Drug Safety Communication - Safety Review of Post-Market Reports of Serious Bleeding Events. Available at: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm282820.htm. Accessed October 29, 2012.

16. Le Heuzey JY, De Ferrari GM, Radzik D, et al. A short-term, randomized, double-blind, parallel-group study to evaluate the efficacy and safety of dronedarone versus amiodarone in patients with persistent atrial fibrillation: the DIONYSOS study. *J Cardiovasc Electrophysiol* 2010;21:597-605.

17. Said SM, Esperer HD, Kluba K, et al. Efficacy and safety profile of dronedarone in clinical practice. Results of the Magdeburg Dronedarone Registry (MADRE study). *Int J Cardiol* 2012 Jul 9. [Epub ahead of print]

18. Chatterjee S, Ghosh J, Lichstein E, et al. Meta-analysis of cardiovascular outcomes with dronedarone in patients with atrial fibrillation or heart failure. *Am J Cardiol* 2012;110:607-13.

19. U.S. Food and Drug Administration. Multaq (dronedarone): Drug Safety Communication - Increased Risk of Death or Serious Cardiovascular Events. Available at: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm264204.htm. Accessed October 29, 2012.

20. Ducharme F, di Salvio F. Anti-leukotriene agents compared to inhaled corticosteroids in the management of recurrent and/or chronic asthma in adults and children. *Cochrane Database of Systematic Reviews* 2004, Issue 1. Art. No.: CD002314.

21. Blais L, Kettani F-Z, Lemiere C, et al. Inhaled corticosteroids vs. leukotriene-receptor antagonists and asthma exacerbations in children. *Respiratory Medicine* 2011;105:846-55.

22. Louie TJ, Miller MA, Mullane KM, et al. Fidaxomicin versus vancomycin for Clostridium difficile infection. *N Engl J Med* 2011;364:422-31.

23. Cornely OA, Crook DW, Esposito R, et al. Fidaxomicin versus vancomycin for infection with Clostridium difficile in Europe, Canada, and the USA: a double-blind, non-inferiority, randomised controlled trial. *Lancet Infect Dis* 2012;12:281-9.

Appendix

U.S. Agency for Healthcare Research & Quality: Methods Guide.

http://effectivehealthcare.ahrq.gov/ehc/products/60/318/MethodsGuide_Prepublication-Draft_20120523.pdf

U.S. Preventive Services Task Force: Methods and Processes.

http://www.uspreventiveservicestaskforce.org/methods.htm

GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924.

http://www.bmj.com/content/336/7650/924