There are so many systems for grading evidence and recommendations out there. Why do we need another one?

Because there is a need for one single system to avoid confusion. The single system should avoid shortcomings of other systems and include their strengths. Some grading systems are based on study design alone without explicit consideration of other important factors in determining quality of evidence. Some systems are excessively complex. An analysis of current grading systems has shown that these and other shortcomings have not been adequately addressed by any one system to date. See: How GRADE compares to other systems

What does the acronym GRADE stand for?
Grading of Recommendations Assessment, Development and Evaluation.

What is the benefit of systematically grading evidence and recommendations?
A systematic approach to grading the strength of management recommendations can minimize bias and aid interpretation of expert-created medical guidelines. Indeed, most guideline groups have accepted the necessity for some sort of grading scheme.

What do you mean with "strength of recommendation"?
Recommendations to administer, or not administer, an intervention, should be based on the tradeoffs between benefits on the one hand, and risks, burden and, potentially, costs on the other. If benefits outweigh risks and burden, experts will recommend that clinicians offer a treatment to typical patients. The uncertainty associated with the tradeoff between the benefits and risks and burdens will determine the strength of recommendations.

Isn't it complicated to have various degrees of recommendations?
It could be. GRADE has only two levels: strong and weak recommendations.

What is considered a strong recommendation?
Based on the available evidence, if clinicians are very certain that benefits do, or do not, outweigh risks and burdens they will make a strong recommendation. Example.

What is considered a weak recommendation?
Based on the available evidence, if clinicians believe that benefits and risks and burdens are finely balanced, or appreciable uncertainty exists about the magnitude of benefits and risks, they must offer a weak recommendation. In addition, clinicians are becoming increasingly aware of the importance of patient values and preferences in clinical decision making. When, across the range of patient values, fully informed patients are liable to make different choices, guideline panels should offer weak recommendations. Example.

What are the factors that influence the strength of recommendation?
There are a number of factors that one needs to consider when grading recommendations. One issue is the confidence in the best estimates of benefit and harm. The rating of methodological quality of the evidence captures that degree of confidence. However, there are a number of other factors that may influence the strength of a recommendation.

What are the factors that determine our confidence in the magnitude of benefits, risks, burden, and costs?
The fundamental study design and additional methodological factors are critically important in determining our confidence in estimates of beneficial and detrimental treatment effects.

What is the fundamental study design difference you are considering?
Because of prognostic differences between groups, and lack of safeguards such as blinding that can avoid biased ascertainment of outcomes, evidence based on observational studies will, in general, be appreciably weaker than evidence from experimental study designs randomized control trials.

In addition to the fundamental study design, what other factors are important to determine the quality of evidence?
Recent years have seen an increased awareness of a number of factors that influence our confidence in our estimates of risk and benefit, such as poor quality of planning and implementation of the available randomized controlled trials suggesting high likelihood of bias; inconsistency of results; indirectness of evidence; and sparse evidence.

How is the quality of evidence categorized in GRADE?
After going through the process of grading evidence, the overall quality will be categorized as high, moderate, low or very low.

The quality of evidence grading sounds rather abstract — e.g., what do you mean with there is "moderate quality evidence"?
We use the following definitions in grading the quality of the evidence: High = further research is very unlikely to change our confidence in the estimate of effect; moderate = further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate; low = further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate; very low = any estimate of effect is very uncertain.

OK, but where do you find all the evidence needed?
Ideally, people who grade evidence should have available to them systematic reviews of the evidence regarding the benefits and risks of the alternative management strategies they are considering.

What if randomized controlled trials are flawed?
Randomized trials with important limitations are categorized as moderate quality evidence. Randomized controlled trials with multiple serious limitations, will fill the low quality evidence category. For more detail, see criteria for assigning grade of evidence.

What are the flaws in randomized controlled trials that decrease the quality of evidence?
Our confidence in the evidence decreases if the available randomized controlled trials suffer from major deficiencies that are likely to result in a biased assessment of the treatment effect. These methodological limitations include a very large loss to follow-up, inadequacy of allocation concealment, or an unblinded study with subjective outcomes highly susceptible to bias. Example.

But what if there is inconsistency and some trials demonstrate a benefit but other do not?
When several randomized controlled trials yield widely differing estimates of treatment effect (heterogeneity or variability in results) investigators look for explanations for that heterogeneity. For instance, drugs may have larger relative effects in sicker, or in less sick, populations. When heterogeneity exists, but investigators fail to identify a plausible explanation, the strength of recommendations from even rigorous randomized controlled trial is weaker. Example.

But what if the evidence from randomized controlled trials is derived from similar, but not identical, populations to those of interest to me?
This should be considered indirect evidence and, to the extent there is uncertainty about the applicability to the relevant population, the strength of evidence will need to be downgraded. Example.

But what if randomized controlled trials included very few patients and observed very few events?
Then again, the quality of the evidence may need to be downgraded. This situation is sometimes called "sparse data". See Example.

What are the circumstances that observational studies can provide moderate or even strong evidence?
While observational studies will generally yield only low quality evidence, there may be unusual circumstances in which this evidence will be classified as moderate or even high quality. For example, on the rare occasions when they yield extremely large and consistent estimates of the magnitude of a treatment effect, we may be confident about the results of observational studies. Example. See also criteria for assigning grade of evidence.
But what if, on rare occasions, all plausible biases from observational studies are working to underestimate an apparent treatment effect?

In other words, the actual treatment effect is very likely to be larger than what the data suggests. Indeed, the quality of this kind of evidence may need to be upgraded. Example.

What to do when the quality of evidence differs across outcomes?

In general, the overall quality of evidence will depend on the lowest quality of all outcomes that are critical for making a decision. Example.

Alright. But what about studies of diagnostic accuracy?

The accuracy of a diagnostic test is a surrogate for important outcomes that might be affected by accurate diagnosis, including improved health outcomes from appropriate treatment and reduced harms from false positive results. However, consideration of the directness of evidence is based on how confident we are of the relation between being classified correctly (as a true positive or negative) or incorrectly (as a false positive or negative) and important consequences of this. The GRADE working group is currently working on further development of a document that will provide further insight into grading the quality of evidence for diagnostic studies. Example.

Where can I read more about the GRADE system?

The 2011 JCE series provides a 20-part guide for systematic review and health technology assessment authors, guideline panelists and methodologists on how to apply the GRADE methodology framework in more detail. Clinicians and policy-making users and anyone new to GRADE may want to consider first reading the 6-part BMJ series from 2008.

3. GRADE guidelines 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011 Jan 3
14. GRADE guidelines 13. Preparing Summary of Findings tables-continuous outcomes. [coming up next in the J Clin Epidemiology]