Clinical guidelines and the question of uncertainty

G. Imberger

Copenhagen Trial Unit, Centre for Clinical Intervention Studies, Rigshospitalet, Copenhagen, Denmark
E-mail: gimberger@gmail.com

Evidence-based medicine aims to make the process of clinical decision making more scientific. Karl Popper once described science as ‘the art of systematic over-simplification’.1 One wonders if evidence-based clinical guidelines sat in Sir Karl’s peripheral vision when he made that comment—no doubt with a wink. With my wink, I write this editorial with the utmost respect for the role of clinical guidelines and for evidence-based medicine in general. It is with this respect that I reflect on the limitations of such evidence, and suggest a structured approach to the consideration of uncertainty when using guidelines. Clinical guidelines can be immensely helpful. However, they are not, and never should be, a didactic instruction for practising medicine.

The GRADE Working Group has developed an approach to grading the quality of evidence and the strength of recommendations when producing clinical guidelines.2 GRADE is a structured and thorough approach, a full description of which is available on their web site (www.gradeworkinggroup.org). While the process is intended for the development of the guidelines, it is a helpful template to understand for anyone who is planning to use those guidelines. Specifically, it is a template that a physician can use to consider the sources of uncertainty when deciding whether to follow a recommendation in a particular clinical situation.

The first important source of uncertainty lies within the definition of the clinical question. A physician needs to assess the applicability of a recommendation to a particular clinical situation: does this recommendation apply to my patient and to my situation? GRADE defines each question (and its resulting recommendation) using PICO: population, intervention, comparison, and outcome.3 Consequently, a recommendation applies to a defined population, having a defined intervention, compared with a defined alternative intervention, and considering a group of defined outcomes. The evidence that has been collected and evaluated has been done so for this particular PICO.

The first variable in the PICO definition is population. No patient will ever be identical to that defined as the population in a recommendation. We are all individuals, after all, further defined by individual situations. As with all applications of evidence-based medicine, external validity can create a practical quandary. The more general the population (in a single trial or in a body of evidence), the more patients to which the evidence may apply. On the other hand, the more general the population, the greater the variation in the patients included, and the less likely that the average patient in the research population is actually similar to your patient lying in the anaesthetic bay. A physician needs to decide if a given patient is sufficiently similar to the population defined in a recommendation. This assessment is not always simple; it combines the relative objectivity of history, examination, and investigation with the more subjective judgements made from clinical experience.

Similarly, the nature of the intervention and comparator need to be sufficiently similar. In many cases, this similarity is straightforward to assess. For example, if Drug A was being compared with placebo in a recommendation, and a physician is deciding between the same two options, then the applicability may be clear. However, an intervention can be a more complex entity, maybe involving particular training, or a particular response from a physician. Moreover, the alternative intervention being considered could be entirely different for a variety of pragmatic and theoretical reasons.

The final variable in the PICO definition is outcomes. The consideration of what outcomes are important will always be gilded with subjectivity. The GRADE process aims to make this consideration as objective as possible, and most importantly, entirely transparent.3 The guideline panel members identify all the potentially relevant outcomes, both desirable and undesirable, for a given PICO. Each panel member then ranks the relative importance of each outcome on a scale of 1–9. If most of the panel ranks an outcome between 7 and 9, that outcome is classified as ‘critical’, between 4 and 7, as ‘important’, and <4, as ‘of low importance’. After the evidence for these outcomes is collected and evaluated, the strength of the recommendation is then based on weighing the desirable outcomes vs the undesirable (considering of course the evidence for how the intervention actually affects these outcomes). If there was substantial variation in the ranking of outcomes within the panel, then it is less likely that a general recommendation is appropriate. Even if there was consensus by panel members, there are many clinical situations where the relative importance of outcomes is different for an individual. Issues regarding resource management can be included as desirable and undesirable outcomes for an intervention, and the benefit/risk balance of these outcomes will also vary in different circumstances.
The GRADE system grades recommendations as strong or weak. A strong recommendation reflects a high degree of certainty that the advantages outweigh the disadvantages and corresponds to saying that most patients should receive this intervention most of the time, and that most informed patients would make the same choice. A weak recommendation reflects more uncertainty about the balance between advantages and disadvantages. This uncertainty may reflect a less clear balance between which outcomes are more important, or it may reflect less confidence in the effect of an intervention on the relevant outcomes.

Confidence in the effect of an intervention comes from the quality of the evidence. The quality, in this instance, is defined as reflecting how confident we can be that the estimates of effect are correct for the relevant outcomes. Of note, the quality assessment may be different for different outcomes, and an assessment of lower quality may not be a reflection that the included trials were poorly designed.

The quality of the evidence is obviously paramount in the process of making recommendations, and it is a substantial source of uncertainty. This uncertainty can come from systematic error, random error, or design error. Systematic error occurs when methodology causes an increase in the risk of bias in the final conclusion. Random error is the ‘play of chance’ and is an inevitable part of inference, representing a permanent barrier to certainty. ‘Design error’ is similar to the assessment of the applicability of a recommendation to a patient, describing the situation where the trials that make up the body of evidence were not designed to investigate the exact PICO of interest. GRADE provides a thorough technique to rate the quality of the evidence for a particular PICO, incorporating the sources of all three types of error.

For a body of evidence to be considered as high quality, you need: a reasonable number of trials with a low risk of bias, evidence that publication bias was unlikely, precision in the overall results, consistency between the results of the trials, and directness. Some situations can improve the quality of the evidence—a large effect or dose–response demonstrated, or if a known confounder is likely to have attenuated the demonstrated effect. Evidence from randomized clinical trials is considered as inherently high quality, and loses quality as any of the above-described elements are assessed as inadequate. Evidence from observational studies is considered inherently lower quality, loses more quality with elements assessed as inadequate, and would need convincing situations that can improve the quality in order to be assessed as higher quality.

If this GRADE system is followed and the assessments are clearly presented, then a physician should be able to clearly read how much uncertainty comes from the quality of the evidence. This uncertainty should be consistent with the strength of the grading. However, as with all research publications, there is a place for critical appraisal by the reader. Evaluation of how the advantages and disadvantages have been weighed up may lead a reader to decide that the uncertainty is greater than that reflected in the strength of the recommendation. Deeper critique may lead that reader to disagree with the quality of the evidence itself. In some cases, such disagreement may rest on subjective judgement. In other cases, there may be more objective flaws in the way the quality of the evidence has been assessed. In both cases, such disagreements increase the uncertainty of the recommendation.

Critical appraisal of guidelines should extend to the techniques used to collect and select the evidence included. Guideline panel members are often experts in the field. Consequently, they are likely to have a strong knowledge of the relevant published evidence and also strong opinions about the effect of interventions. Using their own knowledge of evidence to make a recommendation, without a systematic search and selection process, no matter how experienced or knowledgeable they are, would be an exercise in selection bias, and a source of systematic error. A systematic search, designed to find all of the evidence addressing the defined PICO question, is necessary. Criteria for selection should be pre-specified, objective and reproducible. Achieving both of these tasks with minimal risk of bias can be a time and labour consuming challenge. The task of completing a guideline often, justifiably, means that compromises need to be made; idealism is not a friend to completion. Justifiable as they may be, any short-cuts that had to be made in the process of searching and selecting the eligible trials may increase the uncertainty about whether to follow a corresponding recommendation.

As a summary then, a template on which to consider uncertainty in a recommendation includes three tiers. First, you need to consider whether your patient and clinical situation are sufficiently similar to the PICO defined in the recommendation. Secondly, you need to consider the strength of the recommendation and the uncertainty resulting from the quality of the evidence on which it is based. Thirdly, you need to appraise the methods used to develop the guidelines—the search and selection process, the assessment of quality, and how the advantages and disadvantages were weighed up.

Such a proposal may sound like the unravelling of the advantage of guidelines in the first place. Is it the responsibility of an individual physician to understand and critique the methodology behind recommendations? This responsibility is, to my mind, the same one as understanding and critiquing evidence-based medicine in general. At the very least, a physician should have an appreciation of the limitations. At the other end of the spectrum, anyone suggesting that guidelines be used to benchmark standards of practice should be able to evaluate and clearly defend the amount of uncertainty in the recommendations they propose as those benchmarks.

Anybody who has done a GRADE assessment of the quality of a body of evidence, or even thought about it, will know that it is extremely unusual to be able to tick all the correct boxes. High-quality evidence, by this assessment, is rare. The thing about uncertainty is that most people do not like it. We want to have answers for our patients, clarity in our course of action. An assessment system that clearly highlights the uncertainty can feel like an overly strict school teacher. We throw up our arms and declare that using this system we will
never be able to make a strong recommendation! The other thing with uncertainty, however, is that ignoring it does not diminish its presence. If we want to think of a strong recommendation as one that reflects high confidence in the balance between desirable and undesirable consequences, then we need a system, like GRADE, that assesses confidence accurately.

Where does this leave our terribly uncertain guidelines then? In guideline development, reduction in systematic error is an achievable goal. However, in the evidence itself, systematic error is a large and challenging problem, and random error and the issue of applicability will always remain inevitable sources of uncertainty. With objective reflection, the question of uncertainty should never be about whether it is present. Rather, the question should always be: how much is there? If we apply guidelines as rules by which to practise medicine, they may well fall victim to Sir Karl's cheeky definition of science. If we want guidelines to fulfil the important role of which they are capable, we need to embrace the complexity here, and we need to accept that uncertainty is a dynamic, permanent, and sometimes substantial presence in guidelines.

Declaration of interest
None declared.

References