When Can We Be Confident about Estimates of Treatment Effects?

Moderated by Gordon Guyatt MD

Discussants: Paul Glasziou, MD; Victor Montori, MD; Holger Schünemann, MD

DR. GUYATT: Today, we are going to talk about the concept of quality of evidence.

We need to think about the quality of evidence when we make our treatment decisions and work with our patients to make optimal decisions. We need to have estimates of the benefits or desirable consequences of our treatments, for instance, the prevention of stroke with anti-coagulants, and the downsides or undesirable consequences such as the increased risk of bleeding and the burden associated with anti-coagulant therapy. We may be very confident in those estimates, which makes decision making easier, or we may not at all be confident of those estimates because of inadequate research evidence, and that uncertainty makes it more difficult to select optimal treatment options.

I will illustrate this with a historical example of hormone replacement therapy. Ten or 15 years ago, hormone replacement therapy was widely advocated by the clinical and expert communities, and physicians were strongly encouraged to prescribe hormone replacement therapy for postmenopausal women to reduce cardiovascular risk. We subsequently found out that, certainly for women who are not immediately perimenopausal, hormone replacement therapy does not reduce cardiovascular risk and may even increase it.

The problem with the initial guidelines was that the authors and the practitioners talking to their patients were not sufficiently cognizant of the uncertainty of that evidence. It’s one thing to say to patients, “We believe that hormone replacement therapy lowers your cardiovascular risk;” therefore, we think you should take it.” It’s quite...
Practitioners will certainly want to keep in mind that to make optimal decisions, you need systematic summaries of the best evidence.
``Estimates of Treatment Effects • Guyatt

Everything else, including observational studies, was lower, and that was the core element on which everyone focused. As time went by, we learned about the limitations of the simple hierarchy approach.

Dr. Montori, can you talk about the limitations, and how that has been superseded by a new conceptualization?

DR. MONTORI: Dr. Glasziou just described the proverbial shoulders on which new conceptualizations are now based. This initial notion of a hierarchy of evidence and these manifestations, as Dr. Glasziou has described, were a major breakthrough and one of the key notions that induced a paradigm shift to evidence-based medicine and the way people learned and practiced medicine.

Traditional hierarchies consider study design features to order studies according to the risk involved in introducing an error in the results. According to these hierarchies, randomized trials generate estimates in which we will have greater confidence as compared to the results produced by observational studies. Since these hierarchies were formulated, and thanks to the increasing evidence, we have recognized that not all randomized trials will implement similar degrees of protection against error; therefore, careful considerations of features within each trial are needed to determine how confident we should be in their estimates. Similarly, not all studies measure the outcomes with adequate precision, that is, not all are able to determine the effect of treatment and rule effects that might suggest that a different course of action could be superior. An imprecision in estimates indicates that we are uncertain about where the truth lies, thus reducing our confidence in the estimates of effect.

Another consideration that has emerged is that a study may have provided good evidence or estimates for only one of many outcomes. Yet, other outcomes may be important for making a decision when working with patients or for making broad recommendations.

The other point recognized was that even if 2 studies were similar in terms of their design, they may yield inconsistent results. We needed to examine each of those studies to understand why they led to inconsistent results. However, there was little understanding that inconsistency in itself should lead to a reduction in the confidence that we have about the overall estimates of effect. There’s been increased recognition that the question that the investigators may have been trying to answer could differ substantially from the question decision makers are trying to answer. If there are important differences between trial conditions and outcomes, and what decision makers are interested in, then the treatment effect estimates may not apply, and if they must apply, they do so with limited confidence.

Recognition of error and bias within each of the studies, recognition of consistency across studies, and the extent to which the studies may apply to the situation at hand became important. We are not only required to assess where the hierarchy of evidence of a particular study lies, but rather, what is the overall impact of the body of evidence on our confidence in the estimates. Thus, new hierarchies of evidence focus mostly on the body of evidence and the confidence in the estimates of effect we derive from its totality. New conceptualizations of the hierarchy of evidence have become necessary to account for an explosion of important evidence and because of the need to make decisions based on that evidence at the policy and clinical level.

For predicting risk or prognosis, the cohort study is better than a randomized trial because it involves a more representative group, generating much more valid research findings."
Dr. Glasziou described the origin of evidence hierarchies and Dr. Montori then described how over time, knowledge has changed and how we would look at the overall confidence in estimates of the fact by focusing on more than just original study design.

One of the underlying concerns with these evidence hierarchies has been that over a period of approximately 15 to 20 years, over 100 such systems were developed, and some of these systems were conceptually inappropriate. In 2000, a group of international scientists, practitioners, public health researchers, and individuals from many different disciplines got together and formed the GRADE Working Group. This approach has now been adopted by over 70 organizations, and the group has, in particular, analyzed the factors that make us more confident or less confident in estimating treatment effect or management effect.

The group has also determined factors that influence our confidence when we move from evidence to recommendation; in other words, healthcare recommendations and the overall factors that influence such a movement or this development of recommendation. The group conducted these tasks systematically by exploring all factors that might increase or decrease our confidence. My colleagues have already referred to, what we typically call, risk of bias. This is when studies, although described as a certain study design, still have limitations, and we would lower our confidence in the result of these studies. We've also talked about inconsistency between studies and lack of application of study conditions and outcomes to clinical situations.

There are 2 additional factors (from a total of 5 factors) that would lower our confidence in a body of evidence. These other 2 factors are concerned with the ideal number of patients in the studies or the experience outcomes being very small resulting in lower precision.

The fifth factor that lowers our confidence in the estimates of effect is concerned with studies, for whatever reason, remaining unpublished. In other words, their results remain unknown. This is typically the case when studies do not validate the investigators’ anticipated effects. This publication bias can lead to an important distortion of the overall estimate of effect.

Grading of Recommendations Assessment, Development and Evaluation looks at 5 factors that can lower your confidence in the estimates of treatment effect. This is particularly relevant when we look at the initial study design. There is wide agreement that randomized control trials give us higher confidence when we start looking at evidence because randomization is the best way to protect against bias and confounding. Therefore, in GRADE, randomized control trials are initially classified as high-quality evidence and observational studies, as low or two plus quality of evidence. In GRADE, there are a total of 4 categories of quality of evidence, with randomized control trials starting at the very top, and observation studies, classified into the second of these 4 categories.

If there are no important limitations or reasons to lower our confidence in the estimates of effect in observational studies, then, as Dr. Glasziou described, there may be reasons for why we are more confident in the estimates of effect from observational studies. There are indeed 3 factors in GRADE that allow us to have higher confidence in estimates of effect. I will just mention one of them quickly.

Large treatment effects or management effects generally increase our confidence about an intervention having a positive influence on an outcome—think about the use of insulin in diabetic ketoacidosis. Although there are no randomized control trials, the observations indicate that insulin significantly improves the outcomes of diabetic ketoacidosis, ie, prevents death and complications from diabetic ketoacidosis, and it does so with a very large effect.

There are 5 factors that lower our confidence in the estimate of effect. According to GRADE, the evaluation is done for each of the outcomes determined as important for decision making by a guideline panel. An overall estimate of the confidence in these estimates of effect for a given healthcare question is obtained by GRADE.

As Dr. Montori indicated, one of the big contributions of the GRADE
Estimates of Treatment Effects • Guyatt

Working Group has been to shift the thinking in guideline panels towards outcomes that are important for the patient and defining what outcomes are important, or unimportant for decision making. To sum up, GRADE makes another important contribution by helping define the factors that then influence the development of recommendations. This is based on 4 factors, the first looks at the balance that we benefit from downsizing after having determined what outcomes are important for decision making.

The second is to evaluate how the intervention or management strategy influences the utilization of resources. The third is how important all of the outcomes are relative to each other; for example, a management strategy may reduce mortality, with temporary nausea as the side effect. Under such circumstances, it is likely that people affected by the recommendation would place a higher value on preventing mortality than experiencing temporary nausea.

The fourth factor is how confident we can be in any of the estimates that affect decision making, i.e., moving from the evidence to recommendations. The GRADE approach then provides confidence of recommendations and specific recommendations after evaluating these 4 factors and focusing on what is important to patients. This approach has been used in probably over a thousand recommendations by now and is, as I indicated earlier, widely disseminated.

DR. GUYATT: Thank you very much. I think we’re going to have opportunities to talk about the strength of recommendation issues in another one of these sessions. Let’s focus on the confidence in estimating quality of evidence issues. Does anybody want to comment on what anybody else has said up to now?

DR. MONTORI: One of the things that Dr. Schünemann just pointed out, which I think was also very poorly recognized at the beginning, but is now a really important issue, is the general notion of the corruption of evidence. One of the aspects about confidence of estimate that we haven’t really gotten into is about the studies that are not available in the published record, also called publication bias. What you see published and disseminated beyond the medical journals to the media and public is only a subset of the research that is being conducted. The fact that new approaches, such as GRADE, take that practice into account improves our confidence in the estimates of effect.

This has been critical in areas such as the use of antidepressants, where evidence highlighting the harmful effects of antidepressants was essentially hidden from public view. Estimates that were presented to the Food and Drug Administration were not the same estimates that were published in the medical journals; the latter suggested that antidepressants were more potent. This is the critical element of corruption of the evidence, which is taken into account by approaches such as GRADE. The related element is the issue of fraud, which thankfully, continues to be a smaller problem for which there is, to my knowledge, no specific solution except to promote ethical conduct of research, transparency, and accountability.

DR. GUYATT: Would anyone else like to comment at this point?

DR. GLASZIOU: On Dr. Montori’s point there, I think the major message is that practitioners should be cautious and a little skeptical about any new recommendations that they see because of the difficulties of getting high-quality evidence on the basis of published literature alone and the awareness of problems such as publication bias.

This is even more important for full-time clinicians. If you’re looking at a guideline or a set of recommendations, the basic first question you need to ask is whether the writers have used any sort of hierarchy of evidence in their processing. If they haven’t, then how do they sift through the evidence to find what’s good and what’s bad. If they have used the more traditional approaches, whose history I went through earlier, then that’s a step forward, but it’s still using a pretty primitive sort of tool. The more the
Guyatt • Estimates of Treatment Effects

Clinical Implications

► Quality of evidence refers to confidence that practitioners have in estimates of the treatment effect.

► In the assessment of evidence, there are limitations to just relying on study design, and therefore, it is necessary to take other factors into account.

► When planning to use guidelines or recommendations on treatment effects, the first question to ask is whether the compilers have considered the quality of the studies in determining how confident they should be about the estimates of risks and benefits.

► In order to make optimal treatment or management decisions, practitioners need systematic summaries of the best evidence.

► Several factors affect confidence in the estimates including protection against bias, how directly applicable the findings are to the situation at hand, the consistency of results across studies, and the precision with which the body of evidence estimates the effects of treatment on the most important outcomes to decision makers.
very problematic. The practitioners need summaries of the evidence, as has been emphasized: guidelines are one place to obtain them and systematic reviews are another.

What I wanted to emphasize was that systematic reviews can apply the principles that Dr. Montori laid out in terms of the more sophisticated way of judging confidence in estimates, and the GRADE approach that Dr. Schünemann laid out is perfectly applicable to systematic reviews. In fact, ideally, systematic reviews can be summarized looking at each of the issues that the GRADE Working Group has laid out. Indeed, the world’s leading group focused on systematic reviews, whereas the Cochrane Collaboration has adopted the GRADE approach noting the 5 reasons for rating down confidence in estimates and 3 reasons for rating up that Dr. Schünemann described.

Anybody has any final comments or reflections that you’d like to share with our audience?

DR. MONTORI: Individual clinicians and people making recommendations for clinical policy may experience difficulties in understanding this new conceptualization or applying it to their evidence-based ways of practicing. I think it may be pertinent to point out 3 sources that they could use to follow-up this conversation so that they can improve their understanding. Perhaps, we can all discuss the resources they can use for further learning.

DR. GUYATT: Well, Dr. Montori, we’re running short of time. Having introduced this, why don’t you continue?

DR. MONTORI: Well, I would say for those who want to understand the GRADE system that Dr. Schünemann has pointed out, I believe that currently the best place to find that information is on the GRADE Working Group website. Clinicians using resources such as UpToDate may find instructions on, for example, how to understand the rating of the condition of evidence that those resources are using. Those are a couple of websites that I would direct clinicians to.

DR. GUYATT: Does anybody else wants to comment on the resources issue?

DR. SCHÜNEMANN: Yes, that is a great point. Additional resources with links on the GRADE Working Group website can be found both on the McMaster GRADE Centre website as well as on YouTube, where practitioners might find a series of tutorials that are suited to their needs, starting from brief introductions to very detailed introductions or very detailed descriptions of the GRADE approach.

DR. GLASZIOU: I think it would be great if people could spend time learning how to evaluate and grade evidence themselves. I would probably just warn them that it’s about as difficult as learning how to use a stethoscope, so you have to get a lot of practice and a lot of feedback and guidance to get there. If you don’t feel like you’ve got the time to do that yourself, then you need to find the guidelines or recommendations coming from somebody who has the ability to do it. What I’d advise a clinician to do is either to learn to do it himself/herself or learn to identify resources, guidelines, etc, that have been developed in such a rigorous fashion and have used recommended approaches, such as that of GRADE.

Clinicians need preprocessed information in systematic reviews. Information presented in guidelines is even better for many clinicians because it includes a recommendation and points out where the recommendations are strong and weak. One litmus test for reliability, though not perfect, yet a good indication, is the use of the GRADE approach in summarizing the evidence from a systematic review or in the development of a guideline.

REFERENCES
