Received: 20 October 2017 | Accepted: 6 November 2017

DOI: 10.1111/jep.12857

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

ORIGINAL ARTICLE

Going from evidence to recommendations: Can GRADE get us there?

Mathew Mercuri PhD, Assistant Professor, Doctoral Candidate, Senior Research Associate[1,2,3] | Brian Baigrie PhD, Associate Professor[2] | Ross E.G. Upshur MA MSc MD, Professor[4]

[1] Department of Medicine, Division of Emergency Medicine, McMaster University, Hamilton, Canada

[2] Institute for the History and Philosophy of Science and Technology, University of Toronto, Toronto, Canada

[3] African Centre for Epistemology and Philosophy of Science, University of Johannesburg, Auckland Park, South Africa

[4] Dalla Lana School for Public Health, University of Toronto, Toronto, Canada

Correspondence
Mathew Mercuri, Division of Emergency Medicine, McMaster University, Hamilton General Hospital, 237 Barton Street East, Hamilton, Ontario, Canada, L8L 2X2, McMaster Wing, Rm. 242.
Email: matmercuri@hotmail.com

Abstract

The evidence based medicine movement has championed the need for objective and transparent methods of clinical guideline development. The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) framework was developed for that purpose. Central to this framework is criteria for assessing the quality of evidence from clinical studies and the impact that body of evidence should have on our confidence in the clinical effectiveness of a therapy under examination. Grades of Recommendation, Assessment, Development, and Evaluation has been adopted by a number of professional medical societies and organizations as a means for orienting the development of clinical guidelines. As a result, the method of GRADE has implications on how health care is delivered and patient outcomes. In this paper, we reveal several issues with the underlying logic of GRADE that warrant further discussion. First, the definitions of the "grades of evidence" provided by GRADE, while explicit, are functionally vague. Second, the "criteria for assigning grade of evidence" is seemingly arbitrary and arguably logically incoherent. Finally, the GRADE method is unclear on how to integrate evidence grades with other important factors, such as patient preferences, and trade-offs between costs, benefits, and harms when proposing a clinical practice recommendation. Much of the GRADE method requires judgement on the part of the user, making it unclear as to how the framework reduces bias in recommendations or makes them more transparent—both goals of the programme. It is our view that the issues presented in this paper undermine GRADE's justificatory scheme, thereby limiting the usefulness of GRADE as a tool for developing clinical recommendations.

KEYWORDS

clinical recommendations, evidence-based medicine, GRADE, practice guidelines

1 | INTRODUCTION

The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) framework was created to help healthcare stakeholders make judgements about the quality of clinical research evidence and strength of recommendations for the purpose of developing evidence-based clinical practice guidelines.[1] An evidence hierarchy is the core of this framework. Information derived from randomized trials is rated high, whereas that from observational studies is rated low. The rating can be upgraded if observed effect sizes are large, plausible confounding variables are considered, and/or a dose response is observed. The rating can be downgraded if the estimate of the effect size is imprecise or inconsistent, and/or important threats of bias are not addressed or controlled. A high evidence rating is purported to justify confidence in the estimate of the effect, whereas lower ratings should give pause in that additional research is likely to change that confidence. GRADE also asks users to consider patient preferences and trade-offs with respect to benefits, harms, and costs when making recommendations for clinical practice (eg, clinical practice guidelines). Similar frameworks have been advanced by others (e.g. the Oxford Centre for Evidence Based Medicine "Levels of Evidence").[2,3]

GRADE has been adopted by a number of professional medical societies and organizations entrusted with the delivery of healthcare services.[4] As such, decisions made on the basis of GRADE have direct impact on the organization of healthcare services and patient care. Its influence extends beyond clinical practice guideline development. The World Health Organization, for example, now requires a GRADE like process for the development of their ethics guidelines.[5] Despite its wide acceptance, GRADE has received little by way of critical scrutiny.

Our purpose here is to reveal the logic that sustains the GRADE framework and its conceptualization as a tool for assessing evidence/knowledge. This examination discloses several issues that warrant further discussion. First, the definitions of the "grades of evidence" provided by GRADE, while explicit, are functionally vague; ie, with respect to clinical practice, it is unclear what recommendation goes hand-in-glove with a high or low rating (eg, what is the minimum required rating/level of confidence to integrate findings into practice?). Second, the "criteria for assigning grade of evidence", including both how the level of evidence is assigned and the process of upgrading and downgrading the assignment, are both seemingly arbitrary and arguably logically incoherent. Finally, it is unclear how GRADE users are to integrate evidence grades with other factors that are in play for recommendations for clinical practice, such as patient preferences, and trade-offs between costs, benefits, and harms. The authors of GRADE suggest that accounting for these factors requires judgement, making it unclear as to how the framework reduces bias in recommendations or makes them more transparent—both goals of the programme. It is our view that these issues undermine GRADE's justificatory scheme, thereby limiting the usefulness of GRADE in meeting its stated goals.

2 | GRADES OF EVIDENCE CATEGORIES ARE FUNCTIONALLY VAGUE

Turing first to the definitions of "grades of evidence," the first step for the user is to assess the quality of evidence at hand. This grade "reflects the extent to which confidence in the estimate of the effect is adequate to support recommendations" [[6]; p.995]. The authors claim that an assessment of confidence is important, as "decision makers will be influenced not only by the best estimates of the expected advantages and disadvantages but also by their confidence in these estimates" [[7]; p.924].* The assigned grade is shaped by a number of factors, notably methodological features of the studies that serve as the basis for the evidence. For instance, a study using a randomized controlled trial (RCT) method, barring any major flaws in execution, would serve as the basis for a "high" grade, signifying that "further research is unlikely to change our confidence in the estimate of effect" [[1]; p.1492]. Methodological flaws, including "less" rigorous designs (eg, nonrandomized cohort studies and case studies), or evidence derived from basic sciences (which we take to fall under the "any other evidence" type) lower the grade, thereby the confidence in the estimate of the effect. In sum, methodological features dictate the grade, and the grade determines our level of confidence. The logic here is that our confidence in the estimate of the effect is determined (according to the GRADE framework) by the method of study, within the bounds of the GRADE criteria for assigning a grade of evidence (ie, a modified evidence-based medicine [EBM] hierarchy of evidence).

The stated purpose of GRADE is to introduce a consistent and transparent method for developing clinical practice recommendations based on clinical research. This purpose places a premium on objectivity in the process of developing recommendations, which is consistent with the philosophy of the EBM movement (of which many of its founders and key members are also developers of GRADE). The concept of objectivity is highly controversial in the philosophy of science literature.[8] Within the GRADE framework, objectivity relies on an agreed upon standard for what constitutes evidence of therapeutic effect, and its relative importance in determining what should be recommended in practice. Given the heterogeneity in values among clinicians and their patients, it is unlikely that such a standard exists. Indeed, it is likely that confidence differs between individual practitioners, even where the presented information is the same. A related problem is that our confidence at any moment is predicated on both our prior level of confidence and the nature of the evidence that we are assessing, and not just the methodological features of the study from which that evidence is derived.

Suppose one were interested in evaluating the benefits of a new diuretic for the treatment of hypertension. Ideally, a recommendation to use this agent in clinical practice would depend on our confidence that it is effective at reducing blood pressure (additionally, that there would be no harm to the patient or at least a net positive benefit). The GRADE framework is purported as a tool to ascertain our confidence in the evidence of therapeutic effect at any moment in time. Ideally (in GRADE), our assessment would include high-quality RCTs. Assuming that these trials are reasonably similar with respect to effect size, our confidence in the estimate of the effect should be high, and consequently, the recommendation for the use of the investigated therapy in clinical practice should not be compromised by our confidence in the evidence (although it may still be compromised by the context, eg, affordability and important patient population differences). Now, suppose we did not have access to RCTs. Suppose, instead that the evidence base consists of only a single cohort study. GRADE suggests that our confidence should be low; in short that "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" [[1]; p.1492]. Should we not use such a therapy in clinical practice? Should we use the therapy and expect a wider range of effects in practice (perhaps even a negative effect)? These are important questions for GRADE to address.

*Confidence, as is the case with many of the key terms in GRADE, is not defined. To be confident in something is to have a belief that it is true or can be relied on. In advocating some criteria for what constitutes a confidence level, GRADE is making a strong epistemic commitment with respect to the reliability and validity of a study result (eg, when a therapy will be effective in clinical practice). What is also not clear is the target of the grade—ie, whose confidence? Here, the authors of GRADE seem to assume homogeneity between stakeholder groups, and among individuals within each group with respect to how one comes to believe that something is true or reliable (or perhaps their intention is paternalistic?). Alternatively, one might view this as a normative stance. However, the acceptability of such would require some justification that belief ought to be based on the stated criteria—something GRADE has not achieved. Certainly, a belief is the result of many factors, and there is no expectation that individuals with different values and knowledge bases should be similar in how their beliefs are formed.

The consideration of confidence in this context has a Bayesian flavour. From the Bayesian perspective, the perceived evidentiary value of additional information has an effect on raising or lowering our confidence, or rather, something is considered evidence if it does in fact raise or lower our confidence. However, as is well known, Bayesian formulations are premised upon prior beliefs. These prior beliefs can impact our confidence within the GRADE process, but not in the manner carved out by the authors. It may well be that an individual has a high prior confidence in the effect of a therapy. The impact a prior belief may have on the relationship between evidence and confidence can be illustrated with the following 2 scenarios. Consider a scenario where we have a good description of mechanisms from basic sciences that explain how a therapy works, and/or significant clinical experience with the therapy or a similar class of therapies. In that case (which we will call scenario 1), an observational study demonstrating the expected effect might raise the individual's confidence and may even lead this individual to believe that "further research is unlikely to change confidence in the estimate of effect" [1; p.1492]. GRADE would suggest that it is not appropriate because evidence derived from observational studies should give us "low" confidence. However, this does not ensure that a high level of confidence is not achieved in the user. We can see this point more clearly if we look at the converse situation (which we will call scenario 2). Consider the individual who has an extremely low confidence in a therapy, perhaps due to little experience with that therapy, and no understanding of how it should work (or sparse information about it is available). Should this individual have a high level of confidence after a single large sample, high-quality RCT? GRADE says yes. The fact, however, is that the estimated effect in many studies is often not reproducible.[9-11] We are left in a difficult position: GRADE suggests that one should have little (if any) confidence in a therapy prior to the availability of an RCT,[†] but these considerations point out that confidence is influenced by many factors and cannot be simply reduced to a hierarchy based on GRADE methodology.

The underlying intuition of GRADE (and one with which most would likely agree) is that we should only make recommendations when we have "high" confidence in the estimate of the effect (again, assuming a reasonable harm profile, cost, and no contextual constraints). If not, then what is the point of having criteria and a hierarchy? However, one can imagine that many would express concern over making a strong recommendation based on a single study.[‡] So let us suppose that we avoid that recommendation pending further research. How then do we interpret the first scenario? Is it reasonable to recommend a therapy for practice where we have "low" (GRADE) confidence in the estimate of the effect, and yet equivocate on our recommendation for another even when we have a "high" confidence in the estimate of the effect? The suggestion of criteria for grading confidence should avoid such inconsistencies. More concerning, it is still not clear which level is sufficient for generating a recommendation. As we cannot imagine physicians would be comfortable using a therapy where there is an expectation that the estimate of effect is likely to change, this would seem to preclude any recommendation based on a "low" grade.[§] Does this mean we need "moderate" or better? Should physicians be comfortable with a recommendation that admits further research "may change the estimate"?

As a matter of principle, we could demand that all recommendations start with the highest grade of evidence. However, as we demonstrated earlier, a high grade does not guarantee confidence, and so this principle would be empty, at least from the point of view of GRADE. The authors of GRADE claim that confidence in evidence is only one part of the process and that developing recommendations for practice also requires judgement by the users of the framework. Unfortunately, the GRADE hierarchy offers no mechanism for weighing confidence in that judgement. The problems expressed here are compounded when we examine the criteria for upgrading and downgrading the grade of evidence (which will be discussed in the next section), suggesting issues in the framework independent of whether or not the hierarchy is indeed valid.

3 | CRITERIA FOR UPGRADING/DOWNGRADING CONFIDENCE APPEAR TO BE ARBITRARY

The study type determines the base confidence in the GRADE framework. This relative starting point is based on the EBM hierarchy of evidence. The idea of ranking evidence according to research methods is controversial.[14-16] For example, Borgerson[14] argues that the EBM hierarchy is not epistemically justified on either of the 2 familiar arguments: (1) higher level methods provide special access to identifying causal relationships and (2) evidence derived from higher level methods is relatively less biased. Borgerson also notes that such hierarchies are not widely used in the sciences, generally. This begs the question of what it is about medicine that makes it different than other sciences such that it requires some unique way of "knowing" a relationship between 2 events is causal. A clear defence of the EBM hierarchy (as preferred to alternatives) has yet to be provided.[17] While we share these criticisms, our assessment of the GRADE criteria by which the grade of evidence is upgraded or downgraded suggests issues

[†] The general tendency to dismiss the evidentiary value of non-RCT derived information in clinical practice is a hallmark of the EBM rhetoric, as is reflected in the widely cited quote by David Sackett, "if you find a study was not randomized, we'd suggest that you stop reading it and go on to the next article" [12; p.71].

[‡] Evidence-based medicine advocates highly value systematic reviews of RCTs, presumably because they are cognizant of the fact that any single study is susceptible to "chance" findings (ie, false positive or "Type 1" error, in statistical terms). However, there does not seem to be a higher category of confidence in GRADE above what can be achieved with a single, high-quality RCT. It is doubtful that the authors of GRADE would value a single RCT as equal to a systematic review, as many iterations of the evidence hierarchy advocated by the EBM movement consider systematic reviews are superior to a single RCT (although some would suggest that a large, well designed, and inclusive RCT may in fact supersede a systematic review of small trials.[13] The framework does allow for downgrading the level of evidence, and thus, confidence, on the basis of "sparse" data, which may be the mechanism by which to ensure that systematic reviews do not share the same weight as RCTs in determining our confidence in a therapy or intervention.

[§] One implication of avoiding recommendations based on "low"-grade evidence is that it precludes the use of knowledge where there is a small effect that has been demonstrated using observational study methods. This may be problematic, as such cases are fairly common in medicine. We elaborate on this point in the next section.

in the framework independent of whether or not the hierarchy is indeed valid.

3.1 | Criteria for downgrading confidence

The GRADE framework suggests that our confidence can be reduced in light of specified features of the examined studies, primarily with respect to study execution, reporting, and the estimate of the effect. The framework suggests that the readers decrease the grade 1 level if there is (1) one or more serious limitations to the study quality, (2) an important inconsistency, (3) some uncertainty about directness, (4) imprecise or sparse data, and (5) high probability of reporting bias. The grade can be reduced 2 levels if there is suspicion of a serious limitation to the study quality or there is major uncertainty about directness (ie, "the extent to which the people, interventions, and outcome measures are similar to those of interest" [[1]; p.1491]).

Again, a lack of precision in the language of GRADE is disconcerting. What is a "serious limitation," and how does one distinguish between a "serious" and "very serious" limitation? What is an "important inconsistency"? Such questions are answerable in principle insofar as definitions can be offered or conventions agreed upon by the community, although these added details would require justification, a perhaps laborious task in its own right. What may be more concerning is that even if one were to have a clear and justified definition of these terms, the fact that many of the criteria reduce the grade to the same extent implies that they are all equal in their effect on our "confidence" (ie, one point up or one point down). With this in mind, consider a case where one is developing a recommendation to use a new anti-inflammatory medication to reduce pain. Suppose there is an RCT indicating an anti-inflammatory is more effective than a placebo. We start with a "high" grade of evidence by virtue of the study methodology. Were it the case that the trial was not adequately blinded, one might consider such a serious limitation, resulting in a reduction in the grade to the "moderate" level of evidence (ie, "further research is likely to have an important impact on our confidence in the estimate and may change the estimate" [[1]; p.1492]). Should our confidence be equally reduced in the event that the study was industry sponsored (which raises the probability of reporting bias)? This would also drop the grade by one level (from "high" to "moderate")—suggesting, for example, that the potential for reporting bias should (or does) have the same impact on our confidence as inadequate blinding. No justification is given for why this should be the case.

A further problem is that the effect of a reduction in the grade on our confidence is sensitive to the starting level of evidence. The presented definitions of the grades of evidence suggest an ordinal scale (ie, rank order exists, but the relative difference between ranks is not equal). Thus, moving down one level from high to moderate is not equivalent to moving down from moderate to low. The implication is that a defined methodological bias will have a differential effect on our confidence, depending on if the study is randomized or not.

Another concern is how the criteria handle methodological flaws more generally. It is noncontroversial that limitations in the quality of a study should give pause when interpreting its evidentiary value, irrespective of the content area or its purpose. The GRADE framework reflects this by suggesting the grade of evidence be downgraded. However, according to GRADE, the impact of study quality on our confidence in the estimate of the effect should differ depending on the design. An RCT with very serious limitations to its quality is considered equivalent to a high-quality observational study and superior to clinical experience or a mechanistic model derived from high-quality laboratory studies and biological and physiological principles. Under GRADE, only RCTs can withstand significant concerns over quality, as they will still provide at least some level of preliminary results (ie, "low" grade—further research is likely to have an impact on our confidence). It is not clear why an RCT is somehow more able to overcome flaws in quality. It seems strange that anyone would have any confidence in an estimate of effect based on a flawed study or one of poor quality. The GRADE framework can be salvaged if one considers any study (or body of evidence) of "low" grade to be of no evidence value for clinical practice recommendations. In that case, one could ask why such studies should not be simply disregarded (ie, why provide an evidence grade at all)? The supposition is that such studies can be upgraded if they possess other features, which we will elaborate upon in the next section. Still, it seems counterintuitive to presume that any study can overcome serious limitations in quality—even if the effect size was large (a criterion for increasing the grade), it is odd that one should have any faith that what was observed is in any way meaningful if the methods by which the effect size was acquired are not considered trustworthy.

One final note on the criteria to downgrade—we have discussed the impact of the criteria by using simple cases, ie, a single RCT or a single observational study. The concerns we presented in the foregoing are also relevant when examining a body of evidence, eg, a number of RCTs (in our examples above, one could substitute multiple studies of similar quality and that are in agreement in place of a single RCT or observational study). What is not clear is how the criteria apply when the body of evidence consists of some RCTs of high quality and some with serious limitations to quality. Do a few studies of poor quality warrant a downgrade to "moderate" evidence? Does the existence of one additional study where there is high probability of reporting bias warrant a further downgrade to "low" evidence? It is also not clear what one does in the event that the body of evidence consists of studies of different methods. For example, should the existence of one or more observational studies in a pool of evidence that is predominantly RCTs have any effect on the starting grade? Such issues will require judgement on the part of the user of GRADE, perhaps further undermining the objectivity attributed to a numerical scoring system.

3.2 | Criteria for upgrading confidence

Although high-quality RCTs are considered to have high epistemic value, the GRADE framework does offer a means by which one can salvage information derived from other study methods (in particular, nonrandomized cohort studies), such that our confidence in the estimate of the effect is high.[1,18] The general concern (ie, among EBM advocates) with non-RCT methods is their inability to adequately rule out the effect of all potential confounding variables, known and unknown, on the estimate of the therapeutic effect, hence their starting point as "low" grade evidence. Critics of the RCT[19-21] have argued (effectively in our opinion) that such methods do not guarantee balance between study groups with respect to all potential confounders, and consequently, hold no special privilege in providing evidence of effect. That aside, the GRADE framework does suggest an increase in the grade of evidence (and thus, our confidence) in the event that the association between exposure and outcome is large. The question is how large? Where studies show a relative risk (RR) >2 (<0.5) (ie, "strong evidence of association" [[1]; p.1492]), one may increase the grade one level. A RR >5 (<0.2) (ie, "very strong evidence of association" [[1]; p.1492]) results in an increase of 2 levels. In other words, an observational study with a very strong statistical association can approximate a RCT. There is one caveat: the estimate of the effect must be large, and "based on consistent evidence from two or more observational studies, with no plausible confounders" (for "strong evidence of association), or "based on direct evidence with no major threats to validity" (for very strong evidence of association).[1]

The criterion for upgrading on the basis of the strength of association is problematic. First, no justification is given for why these thresholds were chosen (and are seemingly arbitrary). Second, the criterion stipulates that there be "no plausible confounders" or "major threats to validity".[¶] Setting the aside the absence of criteria in GRADE as to when such circumstances are achieved, it is likely not controversial to assume that any study that manages to rule out plausible threats to bias is generally considered good evidence. We suppose that this is stipulated for non-RCT methods because ruling out threats to bias are thought to be inherent in RCTs—even though this does not hold up to philosophical scrutiny. Still, what is more concerning is that if one were to achieve such a high bar, then why stipulate a minimum effect size? Certainly, there is no such minimum effect size for RCT methods.[#] What then is the added benefit of randomization?

One way to make sense of these thresholds is an underlying concern with contamination of results by unknown confounding variables. As stated above, RCTs are believed to balance the comparison groups on the distribution of these unknown confounding variables, effectively washing out their impact on the estimate of the relative effect of the intervention(s) in question. Let us look at the RR threshold criteria a bit closer. The existence of a threshold suggests that the unknown confounding variables are inflating the effect size. Is this reasonable? Developers of GRADE believe it likely [[6]; p.997]. They might point to empirical data that suggests inflated effect sizes with observational methods compared with RCTs,[23] although other studies show no appreciable difference.[24,25] We are given no reason to assume that confounding factors contribute to the observed effect in only one direction; ie, the unknown confounders only have a positive effect on the outcome. Certainly, it is possible that such factors can drive the effect down so that what is observed is lower than the true effect size (ie, the true effect is masked by confounding variables). Suffice it to say that confounding variables can work in both directions. What makes something an unknown confounder is that we know nothing about its effect on the observed association—to insist that such only inflate the effect size assumes knowledge of the unknown. Furthermore, if unknown confounding variables are in play, it is also possible that they are driving the whole association (ie, the intervention or exposure has no effect on the outcome), in which case no threshold of RR is going to be adequate because no causal relationship between the intervention or exposure and the outcome exists. Without specific knowledge (eg, a theory or empirical evidence) that confounding variables are inflating the effect estimate, the notion of a threshold is a fallacy.

The grading criteria place no value on small effects that are revealed by non-RCT methods; the suggestion is that they are untrustworthy. This scepticism is regrettable simply because some effects in clinical medicine are small. For example, a review of effect sizes among medications for secondary prevention of coronary artery disease showed that only 2 of 21 studies were below the RR threshold of 0.5.[‖][26] It is difficult to ignore a situation where multiple large cohort studies in different contexts or populations demonstrate a small but beneficial effect of an intervention on some health state, after adjustment for known confounders. As there is no mechanism to increase the grade of such evidence, our only option is to hold that additional studies are "likely to change the estimate," giving us low confidence in the estimate of the effect. However, if one were to provide evidence of a dose-response or show that residual confounding variables will reduce the observed effect (both criteria for increasing the grade), then we would be able to move to a moderate level of evidence. Unfortunately, these criteria are also problematic. For one, one cannot assume that all interventions have a monotonic relationship,[**] and there is no mechanism in GRADE to "confidently" capture those that are not without the use of a RCT. Furthermore, although a conservative criterion, as we discussed earlier, there is no reason to believe that the effect of residual confounding variables is unique to non-RCT methods.

4 | THE INTEGRATION PROBLEM-BALANCING CLINICAL TRIAL EVIDENCE WITH OTHER FACTORS

The GRADE framework recognizes that study evidence is not the only factor when considering a recommendation for clinical practice [[30]; p.1049, 1]. After completing the review of evidence and assigning a grade, users are asked to consider the balance of benefits

[¶] Whether this should be based on theory or empirical considerations, GRADE does not specify.

[#] Given that the benefits of many medical therapies/interventions are not without costs (money, time, harms, etc.), it is surprising that consideration of the minimum effect size does not play a more central role in the GRADE framework. Elsewhere the EBM literature raises the concept of the minimally important clinical difference,[22] but this only makes it into GRADE as part of the consideration of trade-offs with harms and costs, although it is not explicitly stated. There is also a noticeable absence (within the GRADE framework) of a process of consultation with patients and payers of healthcare as to what kinds of effects are important and what is the minimum clinical effect from a specified therapy or intervention they are willing to accept.

[‖] The closer the RR is to 1, the smaller the effect. RR < 0.5 (>2) is considered a large effect in the GRADE framework. Three of the studies in the cited review had a RR between 0.9 and 1, suggesting a very small effect.

[**] For example, while there appears to be a linear relationship between ionizing radiation exposure and detrimental health effects, some have argued that there is potential for beneficial effects at very low doses (ie, radiation hormesis).[27,28] Monotonic relationships are also not assumed for the effects of endocrine disruptors in toxicology.[29] If nonmonotonic relationships have potential in risk factor epidemiology (ie, environmental exposure), it is not clear why this should not be the case for therapeutic exposures as well.

with both harms and costs prior to making a recommendation. There is also mention of patient preference, although this is not listed in the 10 step "Sequential process for developing guidelines".[1] Consideration of the balance of benefits, harms, and costs presents 2 problems: (1) no instruction is given on how to measure these factors and (2) it is not clear how one integrates these factors with each other. We will elaborate on both in the following.

When grading evidence, the GRADE framework commits the user to specified criteria for identifying the type of information that warrants belief in an outcome measure (eg, effect size). However, no advice is given regarding measurement. Admittedly, this is an issue for any system of translating study results to clinical practice; this problem is not unique to GRADE. There is a growing body of literature exploring the issue of selecting and measuring outcomes in clinical research.[31] Likewise, how to measure costs (and how to integrate them into healthcare policy) is not straightforward.[32,33] Quantifying harms is particularly challenging, as some are rare and/or are difficult to attribute to a particular source, and many studies are not powered or designed to identify their impact (such harms are often only revealed after a therapy has been released to the market).[34] Furthermore, many of the methods used to best quantify harms, preferences, and costs would constitute low-quality evidence using the GRADE grading system (eg, clinical registries or cohort studies for harms, qualitative or survey methods for preferences, and modelling for costs). The differential attention given to benefits over harms or costs, in part due to the relative ease of measuring benefit, and the fact that the estimate of benefits might be based on more "trustworthy" evidence may both present problems for the process of balancing these factors, potentially skewing the process in favour of overvaluing benefits.

The problem of integration has plagued all aspects of EBM. For example, how to balance patient preferences for a particular therapy, long believed important by advocates of EBM, with evidence of its therapeutic effect as derived from high-quality clinical trials, has received little attention in the core EBM literature.[35] The GRADE framework suffers a similar problem. Benefits, harms, and costs all use different scales, such that integrating these aspects is not a simple process of enumeration. Were they measured on the same scale, one might be able to assess the net effect. However, this would usually take place at a population level (ie, benefits to some and harms to others), which may not reflect management decisions facing physicians when dealing with individual patients in practice—presumably the target of the recommendation.

The authors of GRADE recognize the challenges related to integrating the various factors when generating recommendations for clinical practice. In fact, its authors acknowledge that the framework "does not remove the need for judgment" [[1]; p.1494]. However, the purported value of GRADE is that it can reign in this judgement through a systematic process.[††] How exactly GRADE does this is unclear. As individuals will often differ in how they judge the value of a benefit, harm, or cost, there is no guarantee that different users of GRADE will balance benefits, harms, and costs in the same way. A systematic process can help in this way by providing guidance as to what should be considered and how. Unfortunately, the framework only provides instruction on some aspects of the internal validity of studies and fails to do the same for issues of external validity. The problem of external validity in clinical medicine has been well described elsewhere,[36-38] so we will not discuss it in detail here. Compounding the issue further, even the guidance that is given with respect to internal validity requires judgement on the part of the user (as we described earlier). If judgement by the user group is the ultimate arbiter of clinical value, then it is not clear how the GRADE framework is decidedly different from other methods of recommendation generation that also consider the clinical study results and rely on judgement, such as consensus conferences.[39] Furthermore, there is nothing in the GRADE framework that makes it uniquely transparent, as transparency in any process is not a function of what it judges, but rather how it articulates the basis for the decision.[39,40] Solomon[39] has also noted that it is ironic that the details of the GRADE evidence hierarchy are essentially settled by a consensus of experts, no different than the consensus panels they sought to replace.

5 | DISCUSSION

The GRADE framework is part of an epistemic culture that values information derived from particular study designs (eg, RCT) when determining the evidence value of information relevant to clinical practice. This is apparent in the method of grading evidence. The emphasis of the evidence hierarchy has a number of important repercussions on recommendations for clinical practice. First, high-quality evidence derived from observational studies, laboratory science, or qualitative studies is minimized on account of a strict criteria regarding how to assess bias that is not sensitive to nuances in study design and application. Second, because admittedly important information regarding harms, costs, and preferences often require methods that are considered low on the hierarchy, such information must inevitably be discounted when compared with information regarding effect size when integrating all the important factors to serve as the basis for the recommendation. To do otherwise would create tension within the framework by undermining the notion of confidence one should have in study information put forward by the GRADE working group. To make GRADE an internally consistent framework, one should either eliminate the hierarchy or provide some equivalent method for assessing external validity as is done for internal validity, although the establishment of an analogous hierarchy for harms, preferences, and costs will suffer from the same limitations we present regarding evidence hierarchies.

The GRADE framework is also part of an attempt to align management decisions with the best evidence of benefit. While this is admirable, the success of such an endeavour hinges on the ability of its advocates to identify the best evidence, which in turn requires a clear and defensible definition of what it is to have evidence of something. Epidemiologic study plays an important role in the current clinical

[††] The authors of GRADE suggest that "the GRADE system enables more consistent judgments" so as to "support better-informed choices in health care" [1; p.1490]. The mere suggestion of a "consistent judgement" implies some level of objectivity in the process—different users will more likely come to the same conclusion, whereas with other methods, this is not as likely to be the case. Our analysis suggests that this is something GRADE fails to demonstrate, undermining its value over other methods for recommendation generation.

medicine culture in generating such evidence. The evidence value of such studies is limited by their validity. This is why the grading of evidence is a central focus of the GRADE framework. However, assessing the validity of a study is quite challenging. As Rothman and Greenland note:

> *Although there are no absolute criteria for assessing the validity of scientific evidence, it is still possible to assess the validity of a study. What is required is much more than the application of a list of criteria. Instead, one must apply thorough criticism, with the goal of obtaining a quantified evaluation of the total error that afflicts the study. This type of assessment is not one that can be done easily by someone who lacks the skills and training of a scientist familiar with the subject matter and the scientific methods that were employed. Neither can it be applied readily by judges in court, nor by scientists who either lack the requisite knowledge or who do not take the time to penetrate the work [[41]; p.S150].*

Despite this warning, EBM (and subsequently GRADE) continues to pursue a criteria based method of assessing validity (ie, evidence of effect). Perhaps the GRADE working group recognizes that many clinicians (or other stakeholders, including policy makers, purchasers, and patients) do not have the requisite knowledge to pursue such a task. However, the criteria that are part of the grading of evidence task require that one possess critical appraisal skills to assess methodological features of each study included in the knowledge base. Presumably, if one has the ability to properly assess studies for bias/control of bias, such that our confidence in the effect can be downgraded or upgraded, then the need for criteria is redundant. Elsewhere, a key member of the GRADE working group has advocated for the use of an evidence hierarchy as a heuristic tool to assist physicians in assessing evidence for clinical practice.[42] While the wider community might resign itself to use of a heuristic because of wide variations in critical appraisal skills of physicians, there is no reason why we cannot insist that the process of developing recommendations be limited to those who have the training to properly assess the validity of epidemiologic data, or at least the part of that requiring an assessment of the estimate of effect, harms, preferences, or costs from research studies (which would make the need for a grading system redundant). Eliciting knowledge from scientific study is messy, and application of strict criteria does not make it any cleaner.

ORCID

Mathew Mercuri http://orcid.org/0000-0001-8070-9615

REFERENCES

1. Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group. Grading quality of evidence and strength of recommendations. *Br Med J*. 2004;328:1490-1494.
2. Oxford Centre for Evidence-Based Medicine Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653. Accessed on October 19, 2017.
3. Oxford Centre for Evidence-Based Medicine. Levels of Evidence. 2009. Available at: http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/. Accessed on October 19, 2017.
4. Kavanagh BP. The GRADE system for rating clinical guidelines. *PLoS Med*. 2009;6(9):e10000094
5. World Health Organization. *WHO Handbook for Guideline Development*. 2nd ed. Geneva: WHO Press. Available at: http://apps.who.int/medicinedocs/documents/s22083en/s22083en.pdf; 2014 Accessed on October 19, 2017.
6. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Schunemann HJ. GRADE: what is “quality of evidence” and why is it important to clinicians? *Br Med J*. 2008;336(7651):995-998.
7. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J*. 2008;336(7650):924-926.
8. Reiss J, Sprenger J.. “Scientific objectivity”, The Stanford encyclopedia of philosophy (summer 2017 edition), Edward N. Zalta (ed.); 2017. Available at: https://plato.stanford.edu/entries/scientific-objectivity/. Accessed on October 19, 2017.
9. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015; 116(1):116-126.
10. Ioannidis JPA. How to make more published research true. *PLoS Med*. 2014;11(10):e1001747
11. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124
12. Sackett DL, Rosenberg WMC, Muir Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it Isn't. *Br Med J*. 1996;312(7023):71-72.
13. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. *J Intern Med*. 2003;254(2):105-113.
14. Borgerson K. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspect Biol Med*. 2009;52(2):218-233.
15. Upshur REG. Are all evidence-based practice alike? Problems in ranking of evidence. *Can Med Assoc J*. 2003;169(7):672-673.
16. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet*. 2008;372(9656):2152-2161.
17. Bluhm R, Borgerson K. Evidence-Based based Medicinemedicine. In: Gifford F, ed. *Philosophy of Medicine*. Vol.16 New York: Elsevier; 2011:203-238.
18. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol*. 2011;64(12):1311-1316.
19. Saint-Mont U. Randomization does not help much, comparability does. *PLoS One*. 2015;10(7):e0132102
20. Worrall J. What evidence in evidence-based medicine? *Philos Sci*. 2002;69(S3):S316-S330.
21. Thompson RP. Causality, theories and medicine. In: Illari PM, Russo F, Williamson J, eds. *Causality in the Sciences*. New York: Oxford University Press; 2011:25-44.
22. Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407-415.
23. Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286(7):821-830.
24. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887-1892.
25. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342(25):1878-1886.
26. Caro JJ, Ishak KJ, Caro I, Migliaccio-Walle K, Klittich WS. Comparing medications in a therapeutic area using an NNT model. *Value Health*. 2004;7(5):585-594.
27. Doss M. Linear no-threshold model vs. radiation Hormesis. *Dose-Response*. 2013;11:480-497.

28. Hooker AM, Bhat M, Day TK, et al. The linear no-threshold model does not hold for low-dose ionizing radiation. *Radiat Res*. 2004; 162(4):447-452.

29. Lagarde F, Beausoleil C, Belcher SM, et al. Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment. *Environ Health*. 2015;14(1):13

30. Guyatt GH, Oxman AD, Kunz R, et al. GRADE: going from evidence to recommendations. *Br Med J*. 2008;336(7652):1049-1051.

31. Williamson PR, Altman DG, Blazeby JM, et al. Developing core outcome sets for clinical trials: issues to consider. *Trials*. 2012;13(1):132

32. Drummond MF, Sculpher MJ, Torance GW, O'Brien BJ, Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes*. Third ed. New York: Oxford University Press; 2005.

33. Stone PW, Chapman RH, Sandberg EA, Liljas B, Neumann PJ. Measuring costs in cost-utility analyses: variations in the literature. *Int J Technol Assess Health Care*. 2000;16(1):111-124.

34. Stegenga J. Hollow hunt for harms. *Perspect Sci*. 2016;24(5):481-504.

35. Charles C, Gafni A, Freeman E. The evidence-based medicine model of clinical practice: scientific teaching or belief-based preaching? *J Eval Clin Pract*. 2011;17(4):597-605.

36. Cartwright N, Munro E. The limitations of randomized controlled trials in predicting effectiveness. *J Eval Clin Pract*. 2010;16(2):260-266.

37. Cartwright N. What are randomised controlled trials good for? *Philos Stud*. 2010;147(1):59-70.

38. Horton R. Common sense and figures: the rhetoric of validity in medicine. *Stat Med*. 2000;19(23):3149-3164.

39. Solomon M. *Making Medical Knowledge*. Oxford: Oxford University Press; 2015.

40. Upshur R. Making the grade: assuring trustworthiness in evidence. *Perspect Biol Med*. 2009;52(2):264-275.

41. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health*. 2005;95(S1):S144-S150.

42. Djulbegovic B, Guyatt GH, Ashcroft RE. Epistemological inquiries in evidence-based medicine. *Cancer Control*. 2009;16(2):158-168.

How to cite this article: Mercuri M, Baigrie B, Upshur REG. Going from evidence to recommendations: Can GRADE get us there? *J Eval Clin Pract*. 2018;24:1232–1239. https://doi.org/10.1111/jep.12857