Received: 7 May 2018 | Revised: 15 June 2018 | Accepted: 21 June 2018

DOI: 10.1111/jep.12993

ORIGINAL ARTICLE

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

What confidence should we have in GRADE?

Mathew Mercuri PhD, Assistant Professor, Doctoral Candidate, Senior Research Fellow[1,2,3] | **Brian S. Baigrie PhD, Associate Professor**[2]

[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

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

Abstract

Rationale, Aims, and Objectives: Confidence (or belief) that a therapy is effective is essential to practicing clinical medicine. GRADE, a popular framework for developing clinical recommendations, provides a means for assigning how much confidence one should have in a therapy's effect estimate. One's level of confidence (or “degree of belief”) can also be modelled using Bayes theorem. In this paper, we look through both a GRADE and Bayesian lens to examine how one determines confidence in the effect estimate.

Methods: Philosophical examination.

Results: The GRADE framework uses a criteria-based method to assign a quality of evidence level. The criteria pertain mostly to considerations of methodological rigour, derived from a modified evidence-based medicine evidence hierarchy. The four levels of quality relate to the level of confidence one should have in the effect estimate. The Bayesian framework is not bound by a predetermined set of criteria. Bayes theorem shows how a rational agent adjusts confidence (ie, degree of belief) in the effect estimate on the basis of the available evidence. Such adjustments relate to the principles of incremental confirmation and evidence proportionism. Use of the Bayesian framework reveals some potential pitfalls in GRADE's criteria-based thinking on confidence that are out of step with our intuitions on evidence.

Conclusions: A rational thinker uses all available evidence to formulate beliefs. The GRADE criteria seem to suggest that we discard some of that information when other, more favoured information (eg, derived from clinical trials) is available. The GRADE framework should strive to ensure that the whole evidence base is considered when determining confidence in the effect estimate. The incremental value of such evidence on determining confidence in the effect estimate should be assigned in a manner that is theoretically or empirically justified, such that confidence is proportional to the evidence, both for and against it.

KEYWORDS

evidence-based medicine, philosophy of medicine

1 | INTRODUCTION

The Grades of Recommendations, Assessment, Development and Evaluation (GRADE) framework was developed to assist health care providers in determining recommendations for clinical practice.[1] In a recent keynote address to the International Conference for Evidence-based Health Care Teachers and Developers, Gordon Guyatt, a well-known architect of evidence-based medicine (EBM) and codeveloper of GRADE, noted >100 organizations have now adopted GRADE, including the World Health Organization, the National Health Service, the Province of Ontario, and the European Commission.* As a result of this tremendous market penetration, GRADE has an impact on the training of health care professionals, allocation of health care resources, and ultimately, the management of individual patients' care around the world.

The developers of GRADE (in our opinion, correctly) note that “healthcare workers using clinical practice guidelines and other recommendations need to know how much confidence they can place in the recommendations.”[1(p1490)] Certainly, it is counterproductive to recommend a therapeutic intervention to a patient that one has little belief can be relied on to obtain the desired outcome (ie, where “confidence” in that therapy is low). With that concern in mind, the GRADE framework describes “factors on which our confidence should be based and a systematic approach for making the complex judgements that go into clinical practice guidelines.”[1(p1490)] What precisely should we have confidence about? Here, the authors of GRADE are more explicit: (1) one should “be confident that an estimate of effect is correct,” which they attribute to the “quality of evidence,” and (2) one should “be confident that adherence to the recommendation will do more good than harm,” which they attribute to the “strength of a recommendation.”[1(p1490)] In other words, the quality of evidence gives us confidence in the demonstrated effect of some therapy, and the strength of the recommendation arising from the process (which includes, but is not limited to, assessing confidence in the effect estimate of the therapy) gives us confidence that it will be useful moving forward. In GRADE, confidence in the demonstrated estimate of effect is rated on a scale consisting of four categories (high, moderate, low, and very low). Where confidence falls on this scale is determined through a set of criteria for rating the quality of evidence. These criteria pertain primarily to methodological features of the studies used as the evidence base for determining the estimate of the effect. How one's confidence that adherence to the recommendation will do more good than harm is determined through GRADE is less clear, although some clarification on how the estimate of the effect is supposed to relate to that process, conceptually, has been recently provided.[2]

The Bayesian framework provides another method for assessing confidence, whereby one can use Bayes theorem to adjust his or her level of confidence (or “degree of belief”) on the basis of available evidence and background knowledge. In this paper, we examine the notion of confidence in the GRADE framework through the Bayesian lens. In doing so, this lens shines a light on the rigidity in the GRADE criteria for quality of evidence that we believe is out of step with our intuitions about evidence and its weight in clinical judgement. We will focus much of our attention on how GRADE determines confidence in the observed estimate of effect (ie, the extent to which a therapy can be used to obtain some specified outcome). While much can be said about how GRADE determines confidence in the recommendation, we will avoid too much discussion on that matter because it is less explicit and developed in the framework when compared with how GRADE handles confidence in the estimate of effect.

2 | GRADE'S NOTION OF CONFIDENCE IN THE ESTIMATE OF EFFECT

While the estimate of effect is not the only feature requiring attention when developing a recommendation,[3] it is certainly an important one.[4] One cannot make a recommendation about a therapy if one does not know anything about the magnitude and direction of its effect. In an attempt to clarify what they mean by “certainty of evidence” (or rather, “confidence in effect estimates”; see the next section), the developers of GRADE distinguish between the context of a systematic review, where one seeks “confidence that the estimates of the effect are correct,” and the context of making recommendations, where one seeks “confidence that the estimates of an effect are adequate to support a particular decision or recommendation.”[2(p5)] Again, one would surmise that a “correct” estimate is rather important to garner “adequate support” for a recommendation, and thus, determining our confidence in the estimate of effect from the review of the evidence base (whether this be from a systematic review or selected studies in an evidence panel) is vitally important to the process of arriving at a recommendation for practice.

The GRADE framework offers criteria for determining how confident one should be in the demonstrated effect of a therapy. The criteria pertain to the quality of the evidence base under consideration, in particular the methodological rigour of the studies forming that base, from which the demonstrated effect is derived. The grade is assigned to the evidence base judged to be relevant for the particular effect under consideration. Four levels (or “grades”) of quality are offered, ranging from “high” to “very low.” In the first version of GRADE, a “high” grade indicated that “further research is very unlikely to change our confidence in the estimate of effect.”[1(p1492)] In other words, the evidence base gives one good reason to believe that the estimate of the effect is unlikely to be overturned by further research. A “very low” grade indicates “any estimate of effect is very likely uncertain.”[1(p1492)] The effect on our confidence is not stated, but one might infer that the GRADE developers mean to suggest one should have little or no confidence in an estimate that is very likely uncertain. “Moderate” and “low” grades are supposed to indicate how likely future research will impact on our confidence and the likelihood of a change in the estimate.

A subsequent version of the GRADE framework offered new definitions for the four categories. Now called “quality levels,” the new definitions were to address the issue that in many cases further research may not be forthcoming.[5] Confidence still plays a central role in the definition. For example, an evidence base meriting a “high” quality level rating should make one “very confident that the true effect lies close to that of the estimate of the effect.”[5] That is, the evidence base is thought to warrant a belief that the demonstrated effect is correct or at least close to correct. As we descend the scale, our confidence in the estimate is negatively affected, with growing concern that the true effect may deviate from that observed (ie, may take on a wider range of potential values).

*http://www.ebhcconference.org/previous_editions/2017/presentations/07_02_Guyatt_G.pdf, accessed on 4 May 2018.

We now see that confidence in the effect estimate, in GRADE, is a measure of the quality of the evidence base expressed through a structured quality rating system, or grade of evidence, or something to that effect. How the rating or "grade" (high, moderate, low, and very low) is assigned warrants some description, as it is relevant to our purpose here. GRADE uses a modified EBM "hierarchy of evidence" (eg, Guyatt et al[6]).[†] The initial grade is determined by study design. An evidence base for a specified outcome that consists of one or more high-quality randomized controlled trials (RCTs) is assigned a "high" grade—ie, one should have confidence in the estimate of the effect, that the true effect lies close to the observed effect, and/or further research is unlikely to change our confidence in that estimate. An evidence base derived from observational studies is assigned a "low" grade (our confidence is "limited"). Evidence derived from all other sources, including laboratory sciences, case studies, clinical experience, and mechanical models, is given a "very low" grade (ie, little or no confidence in the estimate). Criteria are provided for increasing or decreasing the grade.[1,5,8] These criteria include study limitations, inconsistency of results, publication bias, imprecision, and indirectness of evidence (factors that decrease quality), large magnitude of effect, dose-response gradient, and plausible confounding (factors that increase quality). The role of the criteria in determining the grade has received criticism, some of which we will touch on in this paper (see Mercuri et al[9]).

The GRADE publications do not explicitly state how the evidence base should be obtained and in particular which studies should be considered relevant. Descriptions of GRADE often refer to a "systematic review" when discussing how the evidence base is acquired.[10‡] We take "systematic" to indicate that the evidence base should be comprehensive and obtained in a transparent and reproducible manner. Whatever the approach, a review of the literature will potentially result in various studies of different design, quality, and relevance. The users of GRADE must determine which of those studies are useful in determining the estimate of the effect. The resulting evidence base might consist of information derived from various study designs. The authors of GRADE provide little advice on how to integrate such studies into the evidence panel. As GRADE is an important part of the EBM movement, one might surmise that users of GRADE will approach determining the relevant evidence base for the recommendation in a similar way to that prescribed by EBM—ie, one should base decisions on the highest "level" of evidence that is available; if, for example, RCTs are not available, then one should look to the next best evidence (eg, observational studies). Presumably, other approaches are available to integrate the evidence from different sources (ie, designs) for the specified outcome, for example, consensus by committee. This process has implications on GRADE's assessment of confidence. If the relevant evidence base consists of RCTs, one can give a starting grade of "high" (ie, very confident) and then proceed to downgrade as appropriate. Likewise, if the relevant evidence base consists of observational studies, one can award a starting grade of "low" (limited confidence) and upgrade or downgrade as appropriate. What does one do in the situation where the deemed relevant evidence base includes both RCTs and observational studies? Do we hedge our confidence rating and assign a "moderate" level pending application of the criteria? GRADE does not provide clear answers to these questions.

3 | CONFIDENCE AS A (SUBJECTIVE) BELIEF IN GRADE AND THE ROLE OF BAYES THEOREM

Earlier we showed how the assessment of the quality of evidence translates to a "grade of evidence" that provides an indication of the level of confidence one should have in the estimate of effect. One can interpret the process of assigning the grade in the framework as conflating quality of evidence with confidence—ie, confidence is a function of one's assessment of the quality of evidence according to the criteria set out in GRADE. Confidence in the estimate of effect and quality of evidence is used interchangeably in the GRADE literature—eg, "the certainty of evidence for those effects (also referred to as quality of evidence or confidence in effect estimates)."[11(p1)] A more recent paper by GRADE's developers more explicitly align confidence and quality of evidence: "GRADE initially referred to 'quality of evidence'; subsequently 'confidence in the estimates' replaced 'quality of evidence'; most recently 'certainty of evidence' has often become the preferred term. These words all refer to the same concept."[2(p5)] Elsewhere, the developers of GRADE admit that "the assessment of evidence quality is a subjective process, and GRADE should not be seen as obviating the need for or minimizing the importance of judgment or as suggesting that quality can be objectively determined."[5(p406)] If our confidence in the effect estimate and the quality of evidence refer to the same thing (as is suggested by the GRADE authors), and if the assessment of the quality of evidence is subjective, then it stands to reason that our assessment of confidence in the GRADE framework is subjective as well. This is not surprising—indeed, the very concept of confidence is subjective.[§]

One could interpret the GRADE view on confidence as a belief, insofar as our confidence in a therapeutic effect is our belief that the true effect is or will approximate that observed in the relevant body of evidence. The authors do suggest the same, "if there are no serious concerns about risk of bias, inconsistency, indirectness, or publication bias, the CI [Confidence Interval] will represent a reasonable estimate of a certainty range, the range of reasonably believable effects of the intervention."[2(p6)] Belief (or confidence) can be measured using a Bayesian framework. Both GRADE and Bayes theorem relate belief in a hypothesis to the evidence for or against its "truth." However, while the Bayesian view reflects our intuitions about evidence, how GRADE measures belief does not, as will become apparent as we work through the scenarios provided later in this paper.

Let us first briefly review the Bayesian framework and how it models belief. Let $P(H)$ indicate the probability a hypothesis H (eg, that

[†] For a critical examination of the EBM hierarchy of evidence, see Borgerson.[7]

[‡] For example, the authors of GRADE state that "the optimal application of GRADE requires systematic reviews of the impact of alternative management approaches on all patient important outcomes."[5(p403)]

[§] The tenth edition of the Concise Oxford English Dictionary[12] defines confidence as a "belief that one can have faith in or rely on someone or something." Belief is "an acceptance that something exists or is true" or a "firmly held opinion or conviction." It is reasonable to interpret belief (and thus, confidence) as personal or taking place within the person's consciousness or perception and, thus, "dependent on the mind for existence" (the definition of subjective).

a therapy is effective to obtain a specified health outcome) is "true." From the Bayesian perspective, $P(H)$ represents one's degree of belief in H given one's background knowledge.[¶] Bayes theorem shows how a rational agent should adjust this degree of belief in H on the basis of the available evidence E using the following equation:

$$P(H|E) = \frac{P(E|H)\times P(H)}{P(E)},$$

where $P(H|E)$ is the probability that H is true given E, $P(E)$ is the probability of obtaining E (independently of H), and $P(E|H)$ is the probability of obtaining E provided H is true (ie, the likelihood of E given H is true). The principle of incremental confirmation posits that E provides some confirmation of H (or rather, increases our belief in the hypothesis) if and only if $P(H|E) > P(H)$. A hypothesis receives greater support from one piece of evidence E over another E' if and only if $P(H|E) > P(H|E')$. Another feature that underwrites the framework is the principle of "evidence proportionism," whereby "a rational believer will proportion her confidence in a hypothesis H to her *total evidence* for H, so that her subjective probability for H reflects the overall balance of her reasons for or against its truth."[13(p10)]

How the Bayesian framework measures degree of belief and how this conception reflects our intuitions of the value of evidence can be illustrated using the following example. Suppose you want to know if Diana plays the piano (H). Having never met Diana you know nothing about her. As few people know how to play piano, you might think it reasonable that the probability that Diana indeed plays the piano is very low (ie, $P(H)$ is very low). You attend a dinner party at Diana's home, where you find a grand piano in her sitting room (E_1). What bearing does this new information have on your belief that Diana plays piano? You note that grand pianos are expensive and quite large, making it unlikely that someone would own one independent of the fact that they play (ie, $P(E_1)$ is quite low). You also reason that someone who owns a piano has more opportunity to play, making it much more likely that they do compared with someone who has less access to a piano. Furthermore, you consider that people who play are more likely to own a piano than people who do not play, as the latter would unlikely want to spend a great deal of money and take up space in their sitting room for something that they do not use. On the basis of the latter two points, your intuition is that it is somewhat likely that you would find a piano in someone's sitting if they played and unlikely if they do not (ie, $P(E_1|H)$ is modest but significant). Putting this information into the equation above, E_1 provides some grounds for belief in H (ie, that Diana plays piano) because the principle of incremental confirmation would hold (ie, $P(H|E_1) > P(H)$). Now suppose that you instead attend a concert where you see Diana perform on the piano (E_2). This piece of evidence should be quite convincing (ie, $P(H|E_2) \approx 1$), as it is very likely that someone performing on the piano could only do so if they know how to play (ie, $P(E_2|H)/P(E_2)$ is high). Intuition would hold that E_2 provides greater support for a belief in H than does E_1—one should be more likely to believe that Diana plays the piano after seeing her perform than from knowledge that she owns a piano. Furthermore, such intuition indeed conforms to principles in the Bayesian framework (ie, $P(H|E_2) > P(H|E_1)$). Knowledge of both E_1 and E_2, conjointly, should raise that belief further. The fact that Diana has red hair (E_3) is likely irrelevant. Finding out that Diana is a trickster might undermine your belief that she plays piano, as the $P(E_1)$ and/or $P(E_2)$ may be quite high (and thus, the ratio of $P(E|H)/P(E)$ is quite low) if you have good reason to suspect that she is appreciably adept at (and has a proclivity for) a piano related ruse.[#]

Let us now look at an example in clinical medicine and contrast the GRADE conception of belief with that derived from the Bayesian framework. We will begin with a hypothetical case. Suppose one is interested if an anticoagulant medication reduces the incidence of stroke (H). A systematic review of the literature yields a single, high-quality multicentre RCT of 2000 participants that shows a 20% relative risk reduction in stroke among those who take the medication compared with those who did not (E_1). The current version of GRADE would consider this "high" quality evidence, and thus, one should be "very confident that the true effect lies close to that of the estimate of the effect."[5(p404)] Does this reflect intuition? If one were to consider that the results of many studies in clinical medicine are not reproducible (E_2), then perhaps one's belief might be tempered somewhat.[14-16] But how much? According to GRADE, this additional consideration should have no effect on our belief.[**] The fact that EBM considers systematic reviews of multiple RCTs to be more reliable in determining the estimate of the effect than a single RCT would suggest that their intuition is that E_2 should have some impact on belief—although the fact that such a consideration was not incorporated into GRADE would suggest E_2 is not enough to push one off a "high" quality rating (ie, there is no level of confidence above what can be achieved with a single RCT, and so a body of evidence consisting of a systematic review of RCTs is effectively the same as one consisting of a single, large, high-quality RCT with respect to the GRADE assessment of confidence in the effect estimate, provided neither body of evidence is appreciably flawed). If so, then what is the purpose of reproducibility or a systematic review? The Bayesian framework allows for E_2 to impact one's belief. This impact may be personal. For example, a sceptical reader of the medical literature might wait for some additional data or justification (ie, more evidence) before committing to a belief

[¶] Bayes theorem allows for many interpretations of the probability function P (objective or subjective). We do not wish to enter a debate about objective vs subjective Bayesian interpretations. Our approach here generally is in step with a subjectivist account of evidence as described by Joyce.[13] This does not mean that our interpretation is strictly subjective in the sense that we believe anything goes provided one's beliefs are coherent. Rather, it is the case that how we set priors and likelihoods is empirically constrained. We believe this subjective interpretation is appropriate given that the GRADE notion of confidence is also subjective (ie, application of the criteria requires judgement by the user of the framework). However, the concern with GRADE we highlight using the Bayesian framework applies regardless of whether one subscribes to an objective vs subjective interpretation.

[#] We thank the anonymous reviewer for pointing out to us the impact of a devious Diana on our belief that she plays piano. Our example shows how one can reason in a Bayesian framework. The same reasoning holds under an objective interpretation, in which case, empirical data may substitute for the assigned probabilities (eg, one can perhaps empirically determine the probability that an individual in a population plays piano and substitute that value in place of our impression that $P(E_1)$ is low).

[**] The GRADE criteria include some nonmethodological or "meta-methodological" considerations (eg, publication bias). However, such considerations are by no means exhaustive and are certainly more restrictive than what is accepted in the Bayesian framework.

aligned with GRADE's definition of high-quality evidence. This skepticism would be reflected in the estimate of the degree of belief within a Bayesian framework, but not in the GRADE framework. The impact of E_2 on belief need not be personal to illustrate concern with the GRADE conception of confidence. For example, one could potentially survey the literature to quantify the probability that published results are overturned when reproduced. If we find that E_2 indeed impacts the estimate of effect, then failure to account for that evidence would result in an inflated confidence in the effect. One might surmise that such evidence is captured in GRADE by the publication bias criterion, although how this is the case, is not clear (nor is it clear how to operationalize the criterion in this way).[††]

Sticking with our hypothetical anticoagulant therapy example, suppose a future review of the literature yields a second, high-quality RCT (E_3), only this one disagrees with the first trial both quantitatively (ie, magnitude of effect) and qualitatively (ie, direction of effect). That is, this second trial suggests that the anticoagulant increases the relative risk of stroke by 20%. What impact should E_3 have on our belief in H? How GRADE deals with this new information is not entirely clear. Under one interpretation of the framework, because E_1 and E_3 are derived from studies of equal methodological rigour, each should independently have the same effect on our confidence in the estimate of the effect; ie, we should expect that the true effect lies close to the estimate in each trial. Using our notation above, one might say that $P(H|E_1) = P(H|E_3)$. This is a bit awkward—how can two pieces of evidence in direct conflict with respect to the estimate of effect each independently meet the threshold for a belief that the effect lies close to the estimate in the trial? Fortunately, the GRADE framework offers a means of reconciling this situation. That is, one could consider the existence of conflicting trials as an "important inconsistency," in which case the quality/grade of evidence would be downgraded one level to "moderate": "we are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different"; or two levels to "low": "our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect."[5(p404)] But which estimate of the effect are we using as our frame of reference? The beneficial effect in the first trial (E_1) or the harmful effect in the second trial (E_3)? The impact of that concern depends on how one interprets the meaning of "substantially different."

What does our intuition tell us? Considering the conflicting evidence, one might be inclined to not hold a belief for or against H. Without knowing anything more about the anticoagulant beyond the two studies, the Bayesian account would suggest that such a belief is not unreasonable. Let us consider how through applying a Bayesian framework one might come to this conclusion. The first step is assigning a value to $P(H)$. One might invoke the notion of clinical equipoise[‡‡] (ie, genuine uncertainty regarding the effect of the anticoagulant on reducing stroke). It is reasonable to assume that E_1 would raise our belief in H, and thus, $P(H|E_1) > P(H)$. It is also reasonable to assume that E_3 would lower our belief in H, and thus, $P(H|E_3) < P(H)$. All things being equal with respect to target population, design, and execution of the studies, the weight we give to each piece of evidence should be roughly the same. Thus, $P(H|E_1\&E_3)$ should approximate $P(H)$, which puts us back to our initial state of clinical equipoise. The preparatory steps (systematic review and preparation of the evidence profile) in the "sequential process for developing guidelines" within GRADE do allow for judgement by the user of the framework.[1] The panel reviewing the literature might use that judgement to determine such a conflicting evidence base as insufficient for a qualified estimate of the effect. Alternatively, the panel might put aside any formal evaluation until which time a meta-analysis (considered a tier above RCT estimates in some EBM evidence hierarchies) can be performed so as to pool the estimates from the discrepant trials. However, that such judgement is allowed in the process would call into question why one should have explicit criteria regarding belief in the first place.

Suppose our literature review yields a mixed bag of evidence from various sources and derived using various methodological approaches. The GRADE framework seems to suggest that the user interprets that evidence as a single unit. This implies that the studies must be integrated in some manner. What is not clear is how one combines estimates from an RCT with those from observational studies, experience, basic sciences, and mechanical explanations. The EBM hierarchy embedded in the criteria would suggest that those studies of higher methodological rigour trump studies lower on that scale. In that case, the existence of a single, large, high-quality RCT should give us good grounds for belief in the effect estimate regardless of what observational studies tell us on the subject.[§§]

Let us look again at our hypothetical anticoagulant and its effect on stroke. We will call the evidence from each RCT E_R, each observational study E_O, and from basic science (including mechanisms) E_S. We found that an evidence base of conflicting RCTs left our confidence in the effect estimate low (or rather, inconclusive). Suppose our evidence base also included two observational studies, E_{O1} and E_{O2}, from different populations showing a 10% and 30% decrease in the relative risk of stroke, respectively, and a series of laboratory studies, E_{Sn}, showing the mechanism by which the anticoagulant will reduce the risk of stroke. The GRADE criteria do not appear to allow for integration of this evidence with the information from the two RCTs we presented

[††] If it was the case that the publication bias criterion captures the issue of reproducibility, this would only show that our ability to pick examples is poor. The issue here is that GRADE is restrictive in how it incorporates evidence into the assessment of confidence, whereas a Bayesian framework is less so, and that such restriction can create tension with our intuition regarding evidence.

[‡‡] Clinical equipoise serves as the ethical basis for clinical trials. While it can be argued that very rarely does one truly know nothing about the effect of a therapy (for example, prior to a trial one may have laboratory data, observational data from clinical practice, or high-quality trial data examining other drugs in the same class as the therapy in a similar population), that it is invoked by the EBM community for the purpose of evidence generation makes it a reasonable (or at least not an unreasonable) place to start when looking at evidence derived from those studies.

[§§] This is not an unreasonable interpretation of the EBM/GRADE view on evidence. For example, Sackett, recognized by many as the grandfather of EBM, points to the "abundant examples of the harm done when clinicians treat patients on the basis of cohort studies,"[17(p177)] in particular the cases of hormone replacement therapy in postmenopausal women[18] and vitamin E in patients with coronary artery disease,[19] where subsequent RCTs showed both to be harmful. A claim of harm in either case (as is suggested by Dr Sackett's words) implies a commitment to the results of the RCT over the observational studies, which revealed the opposite. This suggests that the observational studies should play no role in the formation of one's belief once one or more RCTs are available.

above (or at least how to do so is not clear). One might even interpret statements by key developers of EBM movement (the intellectual movement from which GRADE developed[20]), such as "if you find a study was not randomized, we'd suggest you stop reading it and go on to the next article"[21(p71)] as justification to simply ignore E_{O1}, E_{O2}, and E_{Sn}. Intuition would suggest we do not ignore this potentially valuable information.

Unlike GRADE, the Bayesian framework allows for one to integrate the whole evidence base (ie, the principle of evidence proportionism). That is, one can upgrade or downgrade belief using each piece of evidence, regardless of its methodological underpinnings. In that way, our belief in the effect of the anticoagulant on stroke based on the two RCTs can be moderated by what the two observational studies (E_{O1} and E_{O2}) and the mechanism derived from the laboratory studies (E_{Sn}) demonstrated, even if such information has a lesser incremental evidence value (relative to RCT evidence) due to potential confounding. Consider the case of remote, retroactive intercessory prayer to reduce fever and hospital length of stay among patients with bloodstream infection.[22] As this study used an RCT, the GRADE criteria would suggest that one should have a strong belief that the estimate of the effect (ie, that prayer is beneficial) is close to that shown in the study. Concern about the dearth of data might cause a downgrade to "moderate" quality (belief). However, one might have concern regarding the plausibility of the findings on the basis that a mechanism linking prayer to better outcomes was scientifically implausible given conventional understanding of biology and physics. It is not clear as to how (and if) mechanisms should have any bearing on belief in the effect estimate, as it is not an explicit consideration in the GRADE criteria for grading the quality of evidence.[¶¶] It is safe to say that the medical community has rejected the finding that remote, retroactive intercessory prayer is beneficial in the observed context (ie, for a target population similar to that included in the study sample), the fact that it appeared in the Christmas issue of the *British Medical Journal* (an issue notorious for tongue-in-cheek articles) notwithstanding. We suspect that one would be hard pressed to find an advocate of the EBM or the GRADE framework that would hold the belief that remote, retroactive intercessory prayer is an effective therapy or that one should have even low confidence (in GRADE terms) in the estimate of the effect of that therapy as described in the presented study. If we accept that belief is tied to the principles of evidence proportionism and incremental confirmation as is suggested in the Bayesian framework, the examples above would demonstrate a misalignment between GRADE's stance on what one ought to believe and how beliefs are generated in practice.

4 | DISCUSSION

Belief or confidence that a therapy is effective with some magnitude or that the effect will be within a known range is essential to practicing clinical medicine. Physicians would like to know that what they suggest to patients will work (to some extent), patients would like to know that taking the therapy is worthwhile, and health care managers and insurance companies would like to know that what they are funding is worth the resources. The GRADE framework offers a perspective on how such belief should be generated. Their approach is part of the EBM movement, which seeks to align patient management with those therapies that have been shown effective using justified methods. Rational belief in GRADE conforms to EBM's commitments towards what constitutes methodological rigour for determining therapeutic effect in a clinical population.

The Bayesian framework, on the other hand, assigns belief using the total body of evidence through the principles incremental confirmation and evidence proportionism. An application of the Bayesian framework and these principles revealed a tension between what the GRADE criteria tell us we should believe and intuitive belief that is justified. Such tension is perhaps most apparent when the evidence base consists of studies that use a wide range of methodological approaches and/or of differing levels of rigour. How the GRADE framework allows one to integrate such evidence when determining confidence in the therapeutic effect estimate is not clear or, in some cases, is counterintuitive. A rational thinker uses all the available evidence to formulate a belief. The GRADE criteria, and the hierarchy that underpins it, seem to suggest that we discard some of that information when other, more favoured information is available. To resolve the tension we describe, the GRADE framework should strive to ensure that the whole evidence base is considered when determining confidence in the effect estimate. The incremental value of such evidence on determining confidence in the effect estimate should be assigned in a manner that is theoretically or empirically justified, such that confidence is proportional to the evidence, both for and against it.

ACKNOWLEDGEMENTS

We would like to thank Joseph Berkovitz and Aaron Kenna for their comments on an earlier draft of this manuscript. We would also like to thank the two anonymous reviewers for their suggestions on how to revise this manuscript.

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. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol*. 2017; 87:4-13.
3. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Alonso-Coello PA. GRADE guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength. *J Clin Epidemiol*. 2013;66(7):726-735.

[¶¶] In the 1992 paper presenting EBM as a "new paradigm for medical practice," the Evidence-based Medicine Working Group claimed that "the study and understanding of basic mechanisms of disease are necessary but insufficient guides for clinical practice. The rationales for diagnosis and treatment, which follow from basic pathophysiologic principles, may in fact be incorrect, leading to inaccurate predictions about the performance of diagnostic tests and the efficacy of treatments."[23(p2421)] Evidence-based medicine meant a de-emphasis on pathophysiologic rationale (ie, mechanistic reasoning) and more attention to evidence from clinical research. As a result, mechanistic reasoning would appear at or near the bottom of many EBM evidence hierarchies. Howick,[24] in his examination of the philosophy of EBM, has suggested that there are many cases where acceptance of a therapy in clinical practice was warranted despite a lack of mechanistic evidence (also see Howick et al[25]).

4. Djulbegovic B, Kumar A, Kaufman RM, Tobian A, Guyatt GH. Quality of evidence is a key determinant for making strong GRADE guidelines recommendation. *J Clin Epidemiol*. 2015;68(7):727-732.

5. Balshem H, Helfand M, Schunemmann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-406.

6. Guyatt G, Rennie D, Meade MO, Cook DJ. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Second ed. New York: The McGraw-Hill Companies, Inc.; 2008.

7. Borgerson K. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspect Biol Med*. 2009;52(2):218-233.

8. 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.

9. Mercuri M, Baigrie B, Upshur REG. Going from evidence to recommendations: can GRADE get us there. *J Eval Clin Pract*. 2018;24(5):1232-1239. https://doi.org/10.1111/jep.12857

10. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383-394.

11. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M. GRADE evidence to decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: introduction. *Br Med J*. 2016;353:i2016.

12. *Concise Oxford English Dictionary*. Tenth ed. New York: Oxford University Press Inc.; 2002.

13. Joyce J. Bayes' theorem. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*. Available from https://plato.stanford.edu/archives/win2016/entries/bayes-theorem/. Accessed on May 4, 2018.

14. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015; 116(1):116-126.

15. Ioannidis JPA. How to make more published research true. *PLoS Med*. 2014;11(10):e1001747.

16. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

17. Sackett D. Chapter 6: the principles behind the tactics of performing therapeutic trials. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, eds. *Clinical Epidemiology: How to Do Clinical Practice Research*. Third ed. New York: Lippincott Williams & Wilkins; 2006.

18. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women; principle results from the Women's Health Initiative Randomized Controlled Trial. *JAMA*. 2002;288:32-333.

19. Heart Protection Study Collaborative Group. MRC/BHF heart protection study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomized placebo-controlled trial. *Lancet*. 2002; 36023-36033.

20. Guyatt G. An emerging consensus on grading recommendations? *Chin J Evid Based Med*. 2007;7(1):1-8.

21. 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.

22. Leibovici L. Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *Br Med J*. 2001;323(7327):1450-1451.

23. Guyatt G, Cairns J, Churchill D, et al. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992; 268(17):2420-2425.

24. Howick J. *The Philosophy of Evidence-Based Medicine*. Oxford: Wiley-Blackwell; 2011.

25. Howick J, Glasziou P, Aronson JK. Evidence-based mechanistic reasoning. *J R Soc Med*. 2010;103(11):433-441.

How to cite this article: Mercuri M, Baigrie BS. What confidence should we have in GRADE? *J Eval Clin Pract*. 2018;24: 1240–1246. https://doi.org/10.1111/jep.12993