Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients

Alfonso Iorio,[1, 2] Frederick A Spencer,[2] Maicon Falavigna,[3] Carolina Alba,[4] Eddie Lang,[5] Bernard Burnand,[6] Tom McGinn,[7] Jill Hayden,[8] Katrina Williams,[9] Beverly Shea,[10, 11] Robert Wolff,[12] Ton Kujpers,[13] Pablo Perel,[14] Per Olav Vandvik,[15] Paul Glasziou,[16] Holger Schunemann,[1, 2] Gordon Guyatt[1, 2]

For numbered affiliations see end of article.

Correspondence to:
A Iorio iorioa@mcmaster.ca

Cite this as: *BMJ* 2015;350:h870
doi: 10.1136/bmj.h870

Accepted: 12 December 2014

Introduction

The term prognosis refers to the likelihood of future health outcomes in people with a given disease or health condition or with particular characteristics such as age, sex, or genetic profile. Patients and healthcare providers may be interested in prognosis for several reasons, so prognostic studies may have a variety of purposes,[1–4] including establishing typical prognosis in a broad population, establishing the effect of patients' characteristics on prognosis, and developing a prognostic model (often referred to as a clinical prediction rule) (Table 1).

SUMMARY POINTS

Main concepts

The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) approach defines quality of evidence as confidence in effect estimates; this conceptualization can readily be applied to bodies of evidence estimating the risk of future of events (that is, prognosis) in broadly defined populations

In the field of prognosis, a body of observational evidence (including single arms of randomized controlled trials) begins as high quality evidence

The five domains GRADE considers in rating down confidence in estimates of treatment effect—that is, risk of bias, imprecision, inconsistency, indirectness, and publication bias—as well as the GRADE criteria for rating up quality, also apply to estimates of the risk of future of events from a body of prognostic studies

Applying these concepts to systematic reviews of prognostic studies provides a useful approach to determine confidence in estimates of overall prognosis in broad populations

Lay summary

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach to rating confidence in the results of research studies was initially developed for therapeutic questions

The GRADE approach considers study design (randomized trials versus non-randomized designs), risk of bias, inconsistency, imprecision, indirectness, and publication bias; size and trend in the effect are also considered

Observational studies looking at patients' prognosis may provide robust estimates of the likelihood of undesirable or desirable outcomes in both treated and untreated patients

Patients will often find this information helpful in understanding the likely course of their disease, in planning their future, and in engaging in shared decision making with their healthcare providers

In a previous article, we examined factors that affect confidence in estimates of baseline risk (the risk of bad outcomes in untreated patients), providing examples of how this might influence the confidence in estimates of absolute treatment effect

This paper provides guidance for the use of the GRADE approach to determine confidence in estimates of future events in systematic reviews of prognostic studies in broad categories of patients

Considerations in determining the trustworthiness of estimates of prognosis arising from these types of studies differ. This article covers studies answering questions about the prognosis of a typical patient from a broadly defined population; we will consider prognostic studies assessing risk factors and clinical prediction guides in subsequent papers.

Knowing the likely course of their disease may help patients to come to terms with, and plan for, the future. Knowledge of the risk of adverse outcomes or the likelihood of spontaneous resolution of symptoms is critical in predicting the likely effect of treatment and planning diagnostic investigations.[5] If the probability of facing an adverse outcome is very low or the spontaneous remission of the disease is high ("good prognosis"), the possible absolute benefits of treatment will inevitably be low and serious adverse effects related to treatment or invasive diagnostic tests, even if rare, will loom large in any decision. If instead the probability of an adverse outcome is high ("bad prognosis"), the impact of new diagnostic information or of effective treatment may be large and patients may be ready to accept higher risks of diagnostic investigation and treatment related adverse effects.

Inquiry into the credibility or trustworthiness of prognostic estimates has, to date, largely focused on individual studies of prognosis. Systematic reviews of the highest quality evidence including all the prognostic studies assessing a particular clinical situation are, however, gaining increasing attention, including the Cochrane Collaboration's work (in progress) to define a template for reviews of prognostic studies (http://prognosismethods.cochrane.org/scope-our-work). Trustworthy systematic reviews will not only ensure comprehensive collection, summarization, and critique of the primary studies but will also conduct optimal analyses. Matters that warrant consideration in such analyses include the method used to pool rates and whether analyses account for all the relevant covariates; the literature provides guidance on both questions.[6 7] In this article, we consider how to establish degree of confidence in estimates from such bodies of evidence.

The guidance in this article is directed primarily at researchers conducting systematic reviews of prognostic studies. It will also be useful to anyone interested in prognostic estimates and their associated confidence (including guideline developers) when evaluating a body of evidence (for example, a guideline panel using baseline risk estimates to estimate the absolute effect of

Table 1 | Types and goals of prognostic studies

Study type	Study goal	Examples in field of atrial fibrillation
Overall prognosis	Establish typical risk in broadly defined population*	Risk of bleeding in patients with atrial fibrillation receiving vitamin K antagonists
Prognostic factor	Establish how particular characteristics of patients influence risk	Influence of age on risk of bleeding in patients with atrial fibrillation
Outcome (or risk) prediction model	Development of full prognostic model, simultaneously considering several prognostic factors and classifying patients into various levels of risk	CHADS2 and CHADS-VASC for risk of stroke; HAS-BLED, HEMORRHAGE for risk of bleeding

*It is equally important to estimate likelihood of spontaneous resolution of disease, as discussed in Matthew Thompson et al.[5]

an intervention). People using this article to evaluate an existing systematic review of prognosis may find that the authors did not provide all the necessary information; they may therefore need to consult the primary studies included in the review.

Applying GRADE principles to therapeutic and prognostic questions

The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) approach was developed to facilitate the production of trustworthy clinical practice guidelines. This requires rating the confidence in estimates (quality of evidence) of the effect of interventions on outcomes important to patients and then applying this information to determine the direction and strength of a management recommendation.[8] In brief, the GRADE approach to rating confidence in estimates of a body of evidence for therapeutic interventions firstly considers study design: a body of randomized control trials begins as high quality, whereas a body of observational studies begins as low quality. The approach then involves consideration of five domains that may diminish confidence (rating down)—risk of bias, inconsistency, imprecision, indirectness, and publication bias—and three situations in which confidence might be increased (rating up)—large effect, dose response gradient, and direction of plausible confounding. Depending on the study design and presence of these factors and consequent rating down or rating up, confidence is ultimately designated as high, moderate, low, or very low. Full details of the GRADE approach for evidence about treatments can be found in a series of articles published in the *Journal of Clinical Epidemiology*.[9]

The initial development of the GRADE approach focused exclusively on the effect of alternative management strategies (approaches to treatment, screening, and more recently diagnosis) on outcomes important to patients. GRADE principles may also be applied to prognostic studies answering several questions. One set of questions deals with overall prognosis in a broad population of patients—the topic of this article. Other possible questions include establishing the characteristics of patients (prognostic factors) that, within a population, increase or decrease the risk of an event. A third category of questions is about estimation of individual risk by properly developed and tested clinical prediction rules. Additional details on these three typologies of studies are provided in Table 1 and in a series of recent papers from the PROGRESS working group.[1–4]

In a previous paper, we described how the principles of the GRADE approach might be applied to judging our confidence in estimates of baseline risk needed for appraising the absolute effects of management options on outcomes important to patients.[10] This is essentially the same question that this article covers, framed here as the overall prognosis in a broad group of patients. In this article, we provide detailed guidance for applying criteria to decide on certainty in estimates of prognosis in a broad population of patients. This paper provides specific guidance on the five domains proposed by GRADE as source of limitations that may decrease confidence in estimates of overall prognosis. We will use examples from several systematic reviews to illustrate principles in application of GRADE to bodies of evidence on prognosis in broad populations (Table 2) and describe how to summarize the output of this process in an evidence profile.[11 12] Table 3 presents the GRADE interpretation of its four levels of evidence applied to prognostic studies.

Risk of bias

Considerations of ideal study design

In contrast to questions of treatment, in which the GRADE approach specifies that a body of randomized controlled trials begins as high quality evidence and a body of observational evidence as low quality, in the field of prognosis a body of longitudinal cohort studies initially provides high confidence. Evidence about

Table 2 | Systematic reviews selected to assess application of GRADE to body of prognostic studies

Reference	Disease	Outcome	Studies			Item(s) commented on
			Design	No	No of patients	
Lopes et al[13]	Atrial fibrillation	Bleeding	Any	51	342 699	Risk of bias, inconsistency
Arcelus et al[19]	Eating disorders	Death	Observational	36	17 272	Risk of bias, indirectness
Mohan et al[20]	First stroke	Stroke recurrence	Observational	13	9155	Risk of bias, imprecision, trend
Januel et al[22]	Orthopedic surgery	Venous thromboembolism	Any	47	44 844	Risk of bias, indirectness
Yousef et al[21]	Barrett's esophagus	Cancer	Observational	47	11 279	Risk of bias, inconsistency, publication bias
Su et al[26]	Hemodialysis	Hepatitis C virus infection	Observational	22	34 060*	Inconsistency, imprecision, indirectness
Oldenburg et al[32]	Mild hemophilia	Inhibitor	Observational	4	912	Large effect

*Person years; number of patients not reported.

Table 3 | Definitions of levels of evidence for typical risk of broadly defined population

Quality level	Definition
High	We are very confident that the true prognosis (probability of future events) lies close to that of the estimate*
Moderate	We are moderately confident that the true prognosis (probability of future events) is likely to be close to the estimate, but there is a possibility that it is substantially different
Low	Our confidence in the estimate is limited: the true prognosis (probability of future events) may be substantially different from the estimate
Very low	We have very little confidence in the estimate: the true prognosis (probability of future events) is likely to be substantially different from the estimate

*Prognostic studies measure incidence—that is, target events over time in a population of interest at risk for the target event. The target event can be an adverse outcome (such as mortality) in patients with the disease of interest (for example, recent onset of atrial fibrillation) or the onset of a disease of interest (such as gastric ulcer) in a previously unaffected population. It can also be time to the spontaneous disappearance of a symptom (cough in upper airways or earache in children). Studies assessing prevalence—that is, the number of affected cases in the population of interest—although investigating a similar and complementary topic, are generally cross sectional and will be covered in a separate paper.

prognosis may originate from single arms of randomized control trials, as these could be conceptualized as two single arm observational studies (one being the intervention group, the other control group). When no comparison is being made—that is, when rates measured in one or the other arm, rather than the comparison, is the matter of interest—the distinction between the two designs loses much of its relevance.

In general, however, we will be more confident of estimates of prognosis from observational studies than from randomized controlled trials. The reason for the higher confidence in observational studies is that eligibility criteria for randomized trials usually include filters (for example, restrictions in age, comorbidity, drug intolerance) that exclude patients who are relevant to the prognostic question of interest. Moreover, eligible patients may decline to participate in a randomized trial, and their reasons for declining may be related to their prognosis.

An exception to the general rule that randomized controlled trials are less trustworthy sources of evidence for prognosis are large, simple, pragmatic trials with broad eligibility criteria that may enroll typical patient populations leading to trustworthy estimates of prognosis. For example, consider a systematic review of the frequency of bleeding complications in patients treated with vitamin K antagonists.[13] The authors found that risk estimates from small and middle size randomized controlled trials were lower (1.80 (interquartile range 1.36–2.50) per patient year) than those from large observational studies (median 2.68 (1.75–4.40)) and large randomized controlled trials (3.09 (2.20–3.36)). Risk estimates from observational studies and large pragmatic randomized controlled trials were very similar, suggesting that large pragmatic trials—in contrast to smaller randomized controlled trials—enrolled representative patient populations.

CRITERIA FOR ASSESSMENT OF RISK OF BIAS

Primary

- Was there a representative and well defined sample of patients:
  - Who did not have the outcome of interest at the time of initial observation?
  - Who were at a similar, identifiable, common, and possible early point in the course of the disease?
- Was follow-up sufficiently long and complete?

Secondary

- Were objective and unbiased outcome criteria used?
- Were all characteristics of patients known or suspected to affect the outcome recorded?
- Was there adjustment for important prognostic factors?

Criteria and tools

In the evaluation of risk of bias, we are concerned about limitations in study design and execution of individual studies that may result in overestimates or underestimates of event rates. For example, suboptimal completeness of follow-up may lead to underestimation of incidence rates; misidentification of prevalent cases as incident cases leads to overestimation of incidence. Different instruments for evaluation of risk of bias in prognostic studies exist, including the Quality in Prognosis Study (QUIPS)[14] and a modified version of the Newcastle Ottawa instrument.[15 16] Chapter 13 of the *Cochrane Handbook for Systematic Reviews* provides additional guidance.[17] Although all of these instruments may be useful, they must be tailored to each of the possible specific goals of prognostic studies as discussed above and in Table 1. For example, adjustment for prognostic balance, crucial for establishing the effect of an exposure or intervention, is unimportant in assessing evidence about overall prognosis of broad populations.

In developing our guidance for risk of bias, we focused on the key concepts outlined in the *Users' Guides to the Medical Literature*[18] and summarized in the box. These criteria include the definition and representativeness of the population, completeness of follow-up, and objective and unbiased measurement of outcome. The ultimate goal is to make the quality assessment transparent and structured. As some degree of subjectivity is unavoidable, authors must clearly document the rationale for any decision regarding rating down or up.

Examples from systematic reviews

Arcelus et al evaluated the rate of suicide in patients with eating disorders.[19] Among the 36 included studies, spanning from 1966 to 2010, the diagnostic definition changed substantially over time, making the pooled population unlikely to be representative of patients meeting current diagnostic criteria. Limitations of some studies included unclear reporting of duration of follow-up, possible missed suicides (deaths from suicide not classified as such), and migration of patients from one disease category to another (anorexia, bulimia, associated psychiatric disorders). These serious limitations warrant rating down for risk of bias.

Another systematic review exploring the risk of recurrence after a first stroke, with a follow-up of 10 years,[20] also provides examples of design limitations sufficiently serious to warrant rating down for risk of bias. The definition of stroke recurrence in the individual studies was not always clear, and, even when clearly defined, it differed considerably among the

studies: not all studies distinguished ischemic and hemorrhagic strokes; some included subarachnoid hemorrhage, and some did not. In addition, the studies span 50 years, and changes in the diagnosis and management of stroke likely had an effect on the course of the disease. Also, serious underestimates of the event rate were likely because events occurring within 21–28 days after a first stroke were not accounted for in some of the studies, as the authors thought that distinguishing new events from evolution of the first one was difficult. Therefore, we have lower confidence in estimates of recurrence at the earlier assessments (one month and one year) than for later ones (five and 10 years).

Sensitivity analysis

Investigators assessing the incidence rate of esophageal cancer in patients with Barrett's esophagus,[21] a dysplastic condition that is a precursor to esophageal cancer, found substantial heterogeneity across studies (I^2=66%, 95% confidence interval 55% to 75%). They had, however, made the very reasonable postulate that the rigor of the diagnosis of Barrett's esophagus would explain most of the variability. They found that this was the case: examination of the subsets of studies with (I^2=0) and without (I^2=72%, 60% to 80%) a standardized diagnosis resolved the problem of heterogeneity.

As is the case when judging confidence in estimates in therapy, evaluation of confidence in prognostic estimates presents challenges when risk of bias varies across studies. Inclusion of one or more studies with high risk of bias does not necessarily mean one should rate down confidence in the pooled estimate when, for example, these studies contributed only a small proportion of the events to the pooled event rate.

In addition, risk of bias is just that—a risk that may or may not result in important bias. If investigators show that results are similar in the studies at lower and higher risk of bias, we may reasonably infer that the limitations of the weaker studies did not in fact importantly bias the results.

Sensitivity analyses may, however, suggest that risk of bias is influencing estimates of prognosis. In the example cited above,[21] studies enrolling possibly unrepresentative populations reported a doubling of the estimate of the incidence of cancer in patients with Barrett's esophagus (8.2 (95% confidence interval 5.3 to 12.8) per 1000 patient years), compared with studies enrolling more representative populations (4.9 (3.9 to 6.3) per 1000 patient years).

In another review,[22] studies with and without concerns regarding loss to follow-up produced statistically different estimates of the recurrence rate of venous thromboembolism after total hip replacement (0.31 (0.15 to 0.47) per 100 patients for higher risk of bias studies and 0.91 (0.61 to 1.39) per 100 patients for lower risk of bias studies).

When, as in these two situations, a sensitivity analysis suggests differences in estimates between studies with higher and lower risk of bias, we suggest, in accordance with the standard GRADE approach, using the estimates from the lower risk of bias studies, with no need to rate down confidence for risk of bias.

Inconsistency

The GRADE criteria for judging inconsistency in risk estimates in broad populations, as with other judgments, parallel the criteria for judging inconsistency in risk estimates for the effect of a treatment strategy derived from randomized controlled trials. These include variability in point estimates, extent of overlap in confidence intervals, and where point estimates lie in relation to decision thresholds.

Examples from systematic reviews

Lopes et al,[13] investigating bleeding rates in a large number of observational and randomized trials of patients using vitamin K antagonists, found a more than 10-fold variation in the estimates between studies, for total as well as fatal bleeding, leading to I^2>90%. The absolute rate varied from 0.65 to 7.53 per 100 patient years, a variation that would lead to alternative management approaches. Accordingly, this evidence warranted rating down for inconsistency.

Limitations of testing heterogeneity with I^2

In the study examining the likelihood of stroke recurrence that provided one of our risk of bias examples,[20] seven studies reported events at five years, and four studies reported events at 10 years. At five years, most of the studies still had a large population and provided precise individual estimates for stroke recurrence (fig 1) with a pooled estimate of 26.4% (20.0% to 32.8%) in a random effects meta-analysis; however, heterogeneity, as measured by I^2, was very high (>95%). At 10 years, the individual studies' estimates were somewhat less precise (fig 2), with a pooled estimate of 39.2%

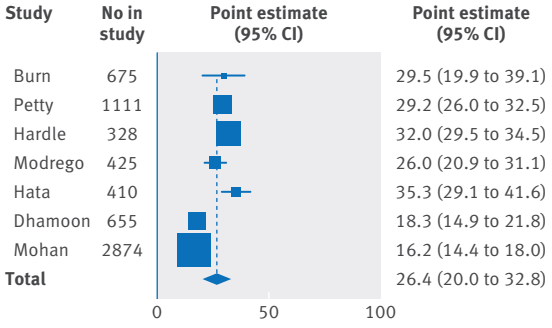

Fig 1 | Risk of recurrent stroke at five years

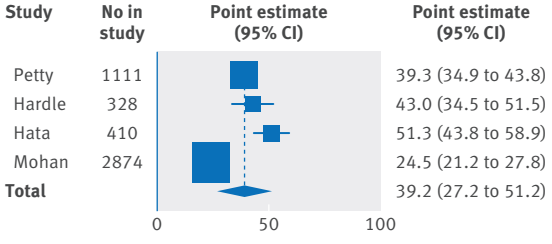

Fig 2 | Risk of recurrent stroke at 10 years

(27.2% to 51.2%), and once again heterogeneity was substantial (I^2>95%).

These results show the challenges of interpreting the I^2 in the context of prognostic studies, where the extremely large sample sizes of the individual studies result in very narrow confidence intervals.[23 24] Under these circumstances, I^2 for the pooled estimate can be extremely high even in the presence of modest inconsistency between studies. When judging inconsistency in such situations, extent of variation in point estimates is far more important; one could even argue that the I^2 is misleading and should not be considered.

Whether the resulting variability in the confidence intervals of the overall pooled estimate is large enough to warrant rating down for inconsistency depends on the effect of the difference on the patient or on the patient's management. For instance, a risk of stroke of 27% at the 10 year follow-up is probably already high enough to warrant high concern from the patient and aggressive management. It seems unlikely that patients' concern or management would increase appreciably only if the risk were truly 51%. Our judgment was that concern and possibly management would not differ across this range, and we therefore did not rate down for inconsistency.

Responding to inconsistency

Authors of systematic reviews should be prepared to face substantial inconsistency in results and therefore generate a priori hypotheses that may explain the heterogeneity they encounter. If the subgroup analyses subsequently show differences across categories (for example, studies of older versus younger patients or more sick versus less sick patients) and meet criteria for credibility—small number of hypotheses specified, a priori direction of observed rates, and consistent biological rationale—chance is a very unlikely explanation.[25] Then one generates separate estimates for the relevant subgroups, potentially resolving the inconsistency problem. This is exactly the same process that we described earlier for dealing with differences in risk of bias, but the hypotheses here relate to clinical rather than methodological differences across studies.

In a review assessing the incidence rate of hepatitis C virus infection in patients undergoing hemodialysis,[26] investigators found considerable inconsistency across studies (1.47 (1.14 to 1.60) per 100 person years; I^2=96.1%; range of estimates in individual studies 0.33–8.05). They had, however, postulated that rates of hepatitis C virus infection in hemodialysis populations would reflect rates in the general populations from which the patients with renal failure arose. They found at meta-regression that their hypothesis accounted for 67.9% of the heterogeneity between the studies. Dividing the studies into those from developed and developing countries and populations with low and high prevalence of hepatitis C virus resolved the inconsistency problem, producing largely non-overlapping estimates of the risk of infection. Figure 3 shows that developed countries can anticipate an incidence ranging from a minimum of 0.16 (if the prevalence of hepatitis C virus positivity is low at baseline) to a maximum of 2.57 (if the prevalence of is high); developing countries had between a seven times higher minimum of 1.07 (low baseline prevalence) and a maximum of 7.19 (high baseline prevalence).

For two reasons, sample size may represent another possible explanation for heterogeneity.[27] Firstly, small sample size is more subject to publication bias because decisions to submit or publish small studies may be driven to a greater extent by the results than is true of large studies. Secondly, small sample size may be a marker for difficult to detect methodological deficiencies that increase bias. These considerations suggest the possible merit of excluding small studies when major differences in results between small and large studies exist. For example, the previously cited systematic review by Yousef et al included 23 small and 24 large studies.[21] The smaller studies resulted in a much higher rate of cancer than did the larger ones (11.6 (8.4 to 16) per 100 patient years versus 4.4 (3.4 to 5.7) per 1000); the latter lower estimate may be more trustworthy. How important these so-called "small study effects" are in the field of prognosis will need further exploration.

Imprecision

Judgments of imprecision of risk estimates are based on the width of the 95% confidence interval around the pooled estimate and the position of the confidence interval relative to a clinical decision threshold. The GRADE rule for prognosis is to rate down confidence in estimates of the event rate if the effect on the patient, or clinical action, would differ depending on whether the upper or the lower boundary of the confidence interval represented the truth.

To illustrate the principle, imagine you are responsible for a cancer program and want to offer intensive follow-up for all patients at a risk higher than 10 per 1000 of developing a cancer. You are considering the previously described systematic review assessing the risk for esophageal cancer in patients with Barrett's esophagus.[21] The estimate of risk crosses the 10 per thousand risk cut-off in men (10.2 (6.3 to 16.4) per 1000 patient years, calculated on the basis of 31 cases and 3445 patient years of observation) thus leaving uncertainty in the decision: if the lower boundary represents the truth you would not follow-up intensively, but if the upper boundary does then you would. Thus, in this situation one would rate down confidence for imprecision.

Among the pre-defined subgroups, you now consider patients with short segment Barrett's esophagus. In this

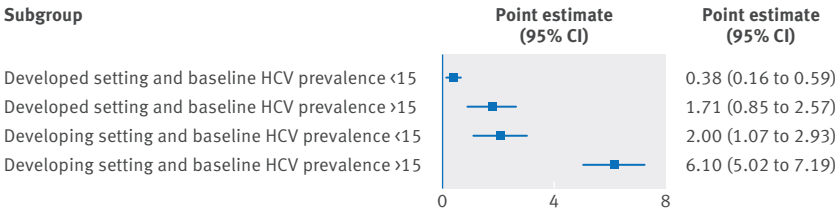

Fig 3 | **Risk of hepatitis C virus infection in hemodialysis patients**

subgroup, the lowest risk of bias studies provide a point estimate of 3.9 per 1000 patient years, with a 95% confidence interval from 3.0 to 5.2, on the basis of 51 cases and 13 677 person years of observation. In this situation, both upper and lower margins of the confidence interval are below your threshold, so you do not need to rate down for imprecision.

It is not, however, in the purview of authors of systematic reviews to make decisions about the appropriate threshold of risk before intensive follow-up is warranted—that should be left to guideline developers. What systematic reviewers can do is point out the implications of the evidence for decision makers. For example, for the Barrett's esophagus screening problem, they could point out that for men, for any clinical management thresholds above 16 per 1000 (the approximate upper boundary of the confidence interval), imprecision is not a problem. For those with short segment Barrett's esophagus, the same is true for any threshold above 5 per 1000.

Indirectness

Authors of systematic reviews need to consider whether the studied population corresponds to their population of interest and whether the measured outcome captures what they believe is important. GRADE refers to these questions, sometimes labeled as generalizability or applicability, as questions of directness.

Consider again the systematic review assessing the frequency of suicide in patients with eating disorders.[19] Imagine a psychiatrist facing the parents of a young woman with uncomplicated anorexia not taking any medications, who are concerned about the likelihood of suicide in their daughter. Using the systematic review as a source of evidence, the psychiatrist would find that most of the patients included in the source studies were diagnosed as having both eating disorders and other psychiatric disorders, were taking psychoactive medication, or reported drug addiction. The psychiatrist might reasonably presume that suicide rates would be different in uncomplicated patients and would therefore be reluctant to apply the results to her patient. Had the systematic review of suicide in patients with eating disorder provided separate estimates for both uncomplicated and complicated patients, and had those estimates differed substantially, the psychiatrist would have relied on estimates from the former population with no need to reduce confidence because of indirectness.

Another example of possible indirectness of population is found in a study of rates of venous thromboembolism after hip or knee surgery.[22] Here, the authors focus only on in-hospital events. The results provide only indirect evidence of the overall risk, including events happening after discharge.

The review of hepatitis C infection in dialysis patients provides an example of indirectness related to outcome.[26] All included studies were conducted before 2006, when testing was based on enzyme-linked immunosorbant assay (ELISA); ELISA is reasonably accurate, but in the setting of hemodialysis direct viral DNA testing by polymerase chain reaction (PCR) is much more sensitive and less affected by hemodilution.[28 29] The viral rates reported in the study provide only indirect evidence regarding the true incidence of infection—presumably this is greater, as we now know that ELISAs can result in a significant number of false negatives, but how much greater remains uncertain.

With respect to indirectness, authors of systematic reviews should consider the range of decision makers who will be using their data. For example, the authors of the systematic review on suicide in eating disorders should be aware that their results have limited applicability to patients with eating disorders without important comorbidities and that such patients are an important subpopulation.[19] The authors of the hepatitis C prevalence review should note that the incidence of diagnosed infection would likely be higher with currently available testing and that this will certainly be important to people making decisions on the basis of the evidence.[26]

Publication bias

The systematic review of Yousef et al,[21] already discussed for imprecision, is a good example of possible publication bias; small studies reported higher rates, suggesting the selective publication of "positive" studies. One might, however, imagine situations in which papers reporting unusually high or low rates of events (outlier cohorts, such as surgical cohorts in which patients do particularly well) are more likely to get published.

Most commonly used statistical tests (for example, Egger's test[30]) for publication bias are applicable when heterogeneity is low and data are normally distributed. As proportions reported in observational studies usually have an asymmetric distribution and inconsistency of results is often high, the use of these instruments is usually not appropriate. As an alternative, tests based on ranking (for example, Begg's test,[31]) may be useful.

Rating up confidence

The GRADE criteria for rating up confidence in treatment studies include large effect, a dose-response gradient, and situations in which all plausible confounders or biases would decrease an apparent treatment effect or would create a spurious effect when results suggest no effect. Analogous situations might exist for the first two of these in prognosis systematic reviews.

An increase in events over time following a well defined pattern (linear or otherwise) would increase confidence in any one of the data points contributing to the linear pattern. Figure 4 depicts such a situation in estimates of stroke recurrence from Mohan et al[20]; only one study directly contributed to the estimates at two years, but our confidence in the estimates increases because they all fall very close to the trend line calculated from the entire dataset.

An example of rating up for large effect comes from the body of evidence assessing the risk of developing inhibitors to therapeutic factor VIII used to prevent bleeding in mild hemophilia,[32] which is associated with

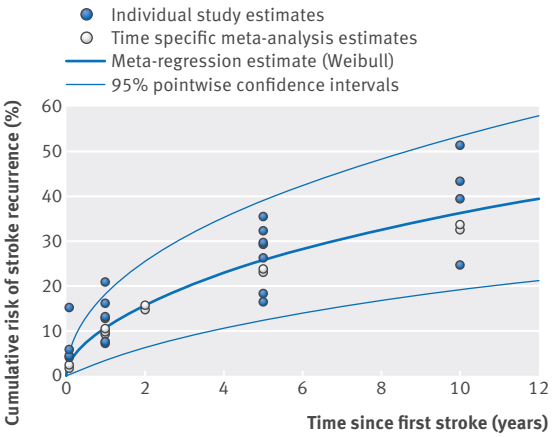

Fig 4 | **Cumulative risk of stroke recurrence after first ever stroke. Reproduced with permission from Mohan et al[20]**

a high risk of bleeding spontaneously into joint, muscles, and vital organs. Data from an international registry indicate that, in patients with specific causative mutations, the risk of developing inhibitors to factor VIII is likely higher than 30% with a lower bound of the estimate as high as 20% (fig 5). If we believe that all or almost all patients with a risk of inhibitor development of substantially greater than 2–5% (the mean risk) would choose an alternative treatment without any risk of inducing inhibitors (such as 1-desamino-8-D-arginine vasopressin), despite lower efficacy and the many inconvenient side effects, we may consider rating up the otherwise very low confidence in our estimates for these patients with specific mutations.

Thus far, we cannot envision a theoretical basis for the third GRADE criterion for rating up related to the nature of plausible biases. That said, as we continue to refine the GRADE approach to prognostic studies, other reasons to rate up confidence in estimates may arise.

Final remarks and future developments

Adopted by more than 70 healthcare organizations, the GRADE approach focuses on evaluating confidence in estimates of the effect of one treatment strategy over an alternative. In this article, we have shown that these same principles work well in assessing bodies of evidence regarding overall prognosis in broad groups of patients. Challenges in defining criteria and providing guidance do, however, remain.

One of these challenges is the gray area between risk of bias and indirectness of evidence. In our suggested framework, extrapolating the estimate obtained in a representative population to a different population or to a specific subgroup is a matter of indirectness, whereas recruiting a non-representative population falls in the risk of bias domain. In other words, when a study sample is representative of the underlying population defined in the study question but does not match the population of interest for a specific clinical question posed by a systematic review author or guideline developer, the question is one of indirectness. When the sample is not representative of the underlying population, because of either limitations in the selection process of an observational study or the patient selection involved in a randomized controlled trial, the relevant domain is risk of bias. Although we have tried to clarify this, gray areas may exist in which the distinction raises a challenging matter of judgment.

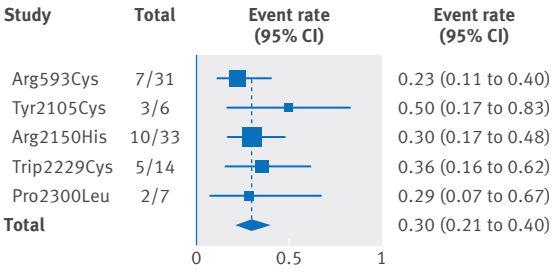

Fig 5 | **Inhibitor risk in mild hemophilia**

One could argue, however, that regardless of categorization (that is, risk of bias or indirectness) what matters most is the overall assessment of the confidence in the prognostic estimates. In this case, the distinction becomes less crucial. The goal is to ensure that readers understand the nature of the question: providing a transparent rationale for judgments will likely achieve that objective.

A second challenging problem arises in the context of random effects pooled estimates of effect with large differences in effect between studies. Using random effects models results in appropriately wider confidence intervals associated with large between study variation Authors of systematic reviews may firstly rate down for inconsistency and then, because of the wide confidence intervals, rate down for imprecision. The risk is double counting inconsistency and imprecision and consequently attributing an excessively low rating of confidence. Systematic review authors should be alert to the risk of such double counting and remember that judgments of confidence in prognostic estimates should not be done mechanistically. Rather, these judgments should be based on an overall assessment of the confidence in prognostic estimates across the five factors that might decrease confidence.

Despite these challenges, applying the GRADE principles for rating confidence in estimates can provide explicit, transparent, and rigorous standards for determining the strength of inferences from bodies of evidence on prognosis in broadly defined populations.

AUTHOR AFFILIATIONS

[1]Clinical Epidemiology and Biostatistics Department, McMaster University, Hamilton, ON, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]National Institute of Health Technology Assessment, Federal University of Rio Grande do Sul Hospital Moinhos de Vento, Porto Alegre, Brazil

[4]Heart Failure and Transplant Program, Toronto General Hospital, University Health Network, Toronto, ON, Canada

[5]University of Calgary, Department of Emergency Medicine Alberta Health Services, Calgary Zone, AB, Canada

[6]Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, Switzerland

[7]North Shore-LIJ Health System, Hofstra North Shore-LIJ Medical School, Hempstead, NY, USA

[8]Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada

[9]Department of Paediatrics, University of Melbourne; Developmental Medicine, Royal Children's Hospital; Murdoch Childrens Research Institute, Melbourne, Australia

[10]Centre for Practice-Changing Research, Ottawa Hospital Research Institute, University of Ottawa, and Bruyère Research Institute, Ottawa, ON, Canada

[11]Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

[12]Kleijnen Systematic Reviews Ltd, York, UK

[13]Department for Guideline Development, Dutch College of General Practitioners, Utrecht, Netherlands

[14]Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK

[15]Department of Medicine, Innlandet Hospital Trust, Division Gjøvik, Norway

[16]Centre for Research in Evidence-Based Practice, Faculty of Health Sciences, Bond University, Gold Coast, Australia

Contributors: All authors contributed equally to the generation of the research hypothesis, participated in the discussion of its content, and approved the final version of the manuscript. AI, FAS, MF, CA, EL, BB, BS, HS, and GG selected the systematic reviews used as examples and prepared the summary of findings tables used in the process. AI drafted the manuscript. GG is the guarantor.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years was disclosed; no other relationships or activities that could appear to have influenced the submitted work. AI, FAS, MF, CA, EL, BB, JH, BS, PP, POV, PG, HS, and GG are members of the GRADE working group.

1 Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al, for the PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380.
2 Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al, for the PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595.
3 Hingorani A, van der Windt D, Riley R, Abrams K, Moons KG, Steyerberg EW, et al, for the PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;345:e5793.
4 Steyerberg EW, Moons KG, Van Der Windt DA, Hayden JA, Perel P, Schroter S, et al, for the PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.
5 Thompson M, Vodicka TA, Blair PS, Buckley DI, Heneghan C, Hay AD, for the TARGET Programme Team. Duration of symptoms of respiratory tract infections in children: systematic review. *BMJ* 2013;347:f7027.
6 Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Res Synth Methods* 2010;1:2–19.
7 Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *J Clin Epidemiol* 2008;61:41–51.
8 Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al, for the GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
9 Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;64:380–2.
10 Spencer FA, Iorio A, You J, Murad MH, Schünemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. *BMJ* 2012;345:e7401.
11 Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables—binary outcomes. *J Clin Epidemiol* 2013;66:158–72.
12 Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles—continuous outcomes. *J Clin Epidemiol* 2013;66:173–83.
13 Lopes LC, Spencer F a, Neumann I, Ventresca M, Ebrahim S, Zhou Q, et al. Bleeding risk in atrial fibrillation patients taking vitamin K antagonists: systematic review and meta-analysis. *Clin Pharmacol Ther* 2013;94:367–75.
14 Hayden J, van der Windt D, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158:280–6.
15 Busse JW, Guyatt GH. Tool to assess risk of bias in case control studies. http://distillercer.com/wp-content/uploads/2014/02/Tool-to-Assess-Risk-of-Bias-in-Case-Control-Studies-Aug-21_2011.doc.
16 Busse JW, Guyatt GH. Tool to assess risk of bias in cohort studies. http://distillercer.com/wp-content/uploads/2014/02/Tool-to-Assess-Risk-of-Bias-in-Cohort-Studies.doc.
17 Reeves BC, Deeks JJ, Higgins JP, Wells GA, on behalf of the Cochrane Non-Randomised Studies Methods Group. Including non-randomized studies. In: Higgins JP, Green S, eds. Cochrane handbook for systematic reviews of interventions. Cochrane Collaboration, 2010:391–432.
18 Randolph A, Cook DJ, Guyatt GH. Prognosis. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. Users' guides to the medical literature—a manual for evidence-based clinical practice. 2nd ed. JAMAevidence, 2008:509–20.
19 Arcelus J, Mitchell AJ, Wales J, Nielsen S. Mortality rates in patients with anorexia nervosa and other eating disorders: a meta-analysis of 36 studies. *Arch Gen Psychiatry* 2011;68:724–31.
20 Mohan KM, Wolfe CD, Rudd AG, Heuschmann PU, Kolominsky-Rabas PL, Grieve AP. Risk and cumulative risk of stroke recurrence: a systematic review and meta-analysis. *Stroke* 2011;42:1489–94.
21 Yousef F, Cardwell C, Cantwell MM, Galway K, Johnston BT, Murray L. The incidence of esophageal cancer and high-grade dysplasia in Barrett's esophagus: a systematic review and meta-analysis. *Am J Epidemiol* 2008;168:237–49.
22 Januel J-M, Chen G, Riffieux C, Quan H, Douketis JD, Crowther MA, et al, for the IMECCHI Group. Symptomatic in-hospital deep vein thrombosis and pulmonary embolism following hip and knee arthroplasty among patients receiving recommended prophylaxis: a systematic review. *JAMA* 2012;307:294–303.
23 Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al, for the GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol* 2011;64:1294–302.
24 Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
25 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
26 Su Y, Norris JL, Zang C, Peng Z, Wang N. Incidence of hepatitis C virus infection in patients on hemodialysis: a systematic review and meta-analysis. *Hemodial Int* 2013;17:532–41.
27 Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ* 2009;339:b4184.
28 Bukh J, Wantzin P, Krogsgaard K, Knudsen F, Purcell RH, Miller RH. High prevalence of hepatitis C virus (HCV) RNA in dialysis patients: failure of commercially available antibody tests to identify a significant number of patients with HCV infection. *J Infect Dis* 1993;168:1343–8.
29 Couroucé AM, Le Marrec N, Girault A, Ducamp S, Simon N. Anti-hepatitis C virus (anti-HCV) seroconversion in patients undergoing hemodialysis: comparison of second- and third-generation anti-HCV assays. *Transfusion* 1994;34:790–5.
30 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34.
31 Begg C, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;50:1088–91.
32 Oldenburg J, Pavlova A. Genetic risk factors for inhibitors to factors VIII and IX. *Haemophilia* 2006;12:15–22.