**Use of GRADE for assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks**

Author list

Farid Foroutan1,3, Gordon Guyatt1, Victoria Zuk1, Per Olav Vandvik2, Ana Carolina Alba3, Reem Mustafa4, Robin Vernooij5, Ingrid Rodriguez6, Zachary Munn7, Pavel Roshanov8, Richard Riley9, Stefan Schandelmaier1, Ton Kujipers10, Reed Siemieniuk1, Carlos Canelo Avbar5, Holger Schunemann1, Alfonso Iorio1.

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Ontario, Canada
2. Department of Medicine, Innlandet Hospital Trust, Division Gjøvik, Norway
3. Ted Rogers Center for Heart Research, Toronto General Hospital, Ontario, Canada
4. Department of Internal Medicine, The University of Kansas Medical Center, United States
5. Iberoamerican Cochrane Centre, Biomedical Research Institute Sant Pau (IIB Sant Pau), Barcelona, Spain
6. Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, Madrid, Spain
7. Joanna Briggs Institute, Faculty of Health and Medical Sciences, The University of Adelaide, Australia
8. Department of Medicine, McMaster University, Ontario, Canada
9. School of Health and Population Sciences, University of Birmingham, Birmingham, United Kingdom
10. Department of Guideline Development, Dutch College of General Practitioners, Utrecth, Netherlands

Address correspondence and requests for reprints to:

Farid Foroutan

Health Research Methods, Evidence, and Impact

McMaster University

foroutaf@mcmaster.ca

**Manuscript details**

Keywords: GRADE, certainty in evidence, prognosis, prognostic factor, guideline, systematic review, subgroup

Word count: 3656

References: 29

Tables = 2

Figures = 3

**Summary concepts**

* GRADE’s approach defines quality/certainty of evidence as certainty in effect estimates; this conceptualization also applies to bodies of evidence addressing overall prognosis in broadly defined populations.
* One can rate certainty of evidence not considering (non-contextualized) or considering (contextualized) the clinical context.
* Here we report how to apply GRADE to risk estimates of future events (i.e. prognosis) in groups of patients identified by a specific prognostic factor using both non-contextualized and contextualized approaches.
* For questions of prognosis, a body of observational evidence (potentially including patients enrolled in randomized controlled trials) begins as high certainty in the evidence.
* The five domains of GRADE for rating down certainty in the evidence, i.e. risk of bias, imprecision, inconsistency, indirectness and publication bias, as well as the domains for rating up, also apply to estimates of associations between prognostic factors and outcomes.
* Applying these concepts to systematic reviews of prognostic factor(s) provides a useful approach to determine the certainty of evidence regarding estimates of difference in risks captured by a prognostic factor, for both contextualised and non-contextualised situations.

**Abstract**

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach to rating certainty in the results of research studies was initially developed for therapeutic questions. The approach considers: study design, risk of bias; inconsistency; imprecision; indirectness, publication bias; magnitude of effect; and dose-response.

Questions about prognosis bear great relevance for decision-making in health care. Studies of prognosis can inform individuals about their likely outcome: for instance, in patients with a new diagnosis of cancer, are they likely to be alive or dead in five years. Further, prognostic studies can aid decisions in those considering treatment: for instance, is one’s risk high enough to use medication to prevent cardiovascular disease. It follows that health care professionals and patients need to know how confident they can be regarding such prognostic estimates.

We have previously provided guidance for using the GRADE approach to determine certainty in estimates of future events in broad categories of patients (overall prognosis). Prognostic studies may also provide more or less robust estimates of the association between patient characteristics (such as their age, sex, and coexisting illness) and undesirable or desirable outcomes. GRADE’s approach to certainty of the evidence aims to inform clinicians and patients of the trustworthiness of the estimates from systematic reviews of studies addressing such individual prognostic factors.

**1.0 Introduction**

Prognosis studies address the likelihood of future health outcomes in a well-defined clinical population1. Such studies often consider prognostic factors, particular characteristics of patients (e.g., age, biomarkers, genetic profile, co-morbidities) associated with the probability of future health outcomes2.

Most often, patients present a constellation of prognostic factors and, accordingly, health care providers rarely use single prognostic factors in clinical decision making. Nevertheless, we have focused this GRADE guidance regarding quality of evidence on a single prognostic factor for one main reason: within a single study and between studies, the certainty of evidence is likely to differ across prognostic factors. For instance, within a study the confidence intervals may be wide for one prognostic factor and narrow for another. Across studies, results may be consistent for one prognostic variable and show large inconsistency for another. Thus, each prognostic factor requires a separate certainty assessment.

One can think of two broad categories of use of single prognostic factors. The first is in relation to study planning and analysis: stratification of randomization, adjusted analysis, and developing a prognostic model. Following previous GRADE guidance, we refer to these situations as non-contextualized4, which implies that there is no direct clinical action associated with the prognostic factor evidence. The second is in relation to clinical decision-making that we refer to as contextualized, which implies that clinicians and patients need prognostic evidence to inform their decisions. Here, we offer guidance both to those conducting systematic reviews of prognostic factors and those interested, from a clinical perspective, in the certainty of evidence supporting prognostic factor estimates.

**2.0 Applying GRADE principles to questions about prognostic factors**

The GRADE approach involves consideration of eight domains that may affect the certainty in the evidence: risk of bias, inconsistency, indirectness, imprecision, publication bias; increase in certainty: large effect, dose response, and plausible confounding. Depending on the study design and issues relating to these domains, certainty is ultimately designated as high, moderate, low, or very low.

To illustrate our approach, we will use real examples from published systematic reviews (Table 1) and provide an example of summary of findings tables including evidence profile for both non-contextualized and contextualized ratings (supplemental material). Table 2 presents the GRADE interpretation of its four levels of evidence applied to prognostic factor studies. Within the supplemental material, we provide additional considerations in addressing certainty of evidence for individual prognostic factors.

**3.0 Risk of bias**

*3. 1 The ideal study design*

Best evidence regarding prognostic factors usually originates from observational studies (cohort studies, registries, or database linkage studies); such studies start with high certainty ratings3. Although secondary analyses of randomized control trials (RCTs) can also provide evidence regarding prognosis, observational studies typically yield higher certainty because eligibility criteria for RCTs usually include restrictions that exclude patients relevant for assessment of prognostic factors. Moreover, eligible patients may decline to participate in RCTs, and their reasons for declining may be related to their prognosis.

For example, a study that compared differences in characteristics of 4713 enrolled patients in the Euro Heart Survey on Coronary Revascularization to 8647 patients enrolled in 14 major RCTs comparing percutaneous coronary intervention with coronary artery bypass grafting11 reported that patients enrolled in trials were significantly younger, and less frequently suffered from hypertension, hyperlipidemia, diabetes, peripheral vascular disease, and cerebrovascular disease. If investigators do use RCT data to address prognosis, they may use either the control group alone, or the entire RCT cohort, in which case adjustment for the intervention will be required.

*3.2 Assessing risk of bias in individual studies*

When evaluating the risk of bias, we are concerned about elements in study design and conduct that may result in over- or underestimation of the true risk ratio, hazard ratio or odds ratio. The QUality In Prognosis Studies (QUIPS) provides a useful instrument developed for evaluation of risk of bias in studies addressing prognostic factors12. Risk of bias instruments such as Prediction model Risk Of Bias ASsessment Tool (PROBAST) may also provide useful considerations for assessment of risk of bias13. Chapter 13 of the Cochrane Handbook for Systematic Reviews of Interventions provides additional guidance14.

*3.3 Adjusted and unadjusted analyses*

In most instances, prognostic factors are correlated with each other, and as a result their individual associations with outcomes may be potentially misleading. For instance, consider a critical care physician estimating a patient’s risk of serious bleeding as a result of gastric stress ulceration. The clinician will find that – when considered individually – mechanical ventilation, coagulopathy, hepatic failure, sepsis and hypotension all increase the risk of bleeding by fivefold or more15. The clinician facing a patient in whom all these factors exist might conclude that the risk of bleeding is extremely high.

A multivariable adjusted analysis, however, revealed that only the first two of these are independent prognostic factors – the others derive their apparent predictive power from their association with the first two, and clinicians should consider only mechanical ventilation and coagulopathy in assessing risk of bleeding. Thus, when clinicians consider multiple factors simultaneously in making prognostic estimates, adjustment in multivariable analyses including all factors is required to generate useful estimates of added prognostic value. The same is true for clinical investigators using prognostic studies for stratification, adjustment, or prediction guide construction2. Aside from standard multiple regression, ‘independent’ or ‘adjusted’ prognostic value may also be derived from score matching, inverse probability weighting, marginal structural modeling, and machine learning.

The situation differs, however, in a less common scenario in which clinicians rely on a single easily measured prognostic factor that is related to numerous less important prognostic factors, and is almost as good as an overall model including all factors. Consider, for example clinicians assessing patients for the need of diagnostic imaging to rule out venous thromboembolism (VTE). Here individuals with a negative D-dimer alone have an expected 98.9% probability of uneventful follow up over three months 16 17, almost identical to the 98.6% likelihood of not having a clot for patients with a Wells clinical predictive model score (considering 7 prognostic factors) of ≤118. In this case, in patients with a negative D-dimer, the unadjusted estimate provides essentially the same level of information as the model that includes all relevant variables.

This latter situation is, however, unusual. Therefore, in the remainder of this paper, we will consider the more common context in which clinicians simultaneously consider the impact of multiple factors on patients’ prognosis, and so require the adjusted effect of a prognostic factor. In such situations lack of simultaneous consideration of prognostic factors represents a source of bias.

For instance, Sanchez et al. conducted a systematic review evaluating the prognostic association between right ventricular (RV) dysfunction and in-hospital mortality in patients with acute pulmonary embolism19. The authors included seven studies with 666 patients and reported a relative risk of 2.43 (95% CI of 1.33 to 4.45), suggesting that the risk is 2.43 times larger in those with RV dysfunction compared to without. All relative risks being pooled, however, are unadjusted. Clinicians, in addressing the prognosis of patients with pulmonary embolus, will consider factors other than RV dysfunction, including systolic blood pressure, the extent of hypoxemia, and heart and respiratory rates. Therefore, adjusted analysis is crucial to avoid a misleading conclusion for this assessment. Thus, the study’s failure to adjust requires rating down for risk of bias.

The discussion thus far has made evident that studies of prognostic factors should, ideally, conduct a multivariable analysis that includes all prognostic factors associated with the outcome of interest. In most instances, however, the number and choice of factor adjustment will vary across studies. If the literature is dominated by studies that fail to adjust for one or more crucial predictors, adjusted estimates may be at high risk of bias.

*3.4 Results may influence risk of bias judgements*

Prior to rating down for risk of bias, authors of systematic reviews should determine if studies at high risk of bias have actually biased the meta-analysis results. If a body of evidence includes a robust collection of studies at both high and low risk of bias, and a subgroup analysis shows that the high risk of bias studies differ importantly in their estimates from the low risk of bias studies, one should rely only on estimates from the latter. If, however, studies at high and low risk of bias provide similar estimates of association, authors can narrow confidence intervals by including all studies and not rate down for risk of bias. In making this judgement, authors need to consider the weight that each study contributes to the final estimate of effect when considering the impact of including studies at higher risk of bias.

For instance, the review by Cheng et al. included studies evaluating the association between smoking and onset of venous thromboembolism (VTE) in patient at risk for VTE20, and reported a relative risk (RR) of 1.19 (95% CI of 1.15 to 1.22). The authors report that some studies adjusted only for body mass index, others for body mass index, cholesterol, diabetes, hypertension, alcohol consumption, and physical activity. This variation raises the possibility that a more comprehensive adjustment would provide less biased results. Visual inspection of the point estimates and 95% confidence interval, however, shows minimal difference between studies with more or less comprehensive adjustment. In this instance, the similar associations suggest that the extent of adjustment is unimportant, and there is no need to rate down for risk of bias.

In the same review, the authors included studies that utilized different criteria for diagnosis of VTE20. Two of the largest studies following over 3 million individuals utilized physician billing codes for determining cases of VTE, whereas the smaller studies utilized radiologic criteria. The physician billing codes, likely to be less accurate than explicit radiologic criteria, represent a potentially important source of bias. The two high risk of bias studies again provide similar estimates for the association between smoking and VTE as the low risk of bias studies. One can therefore include all studies without rating down for risk of bias.

**4.0 Inconsistency**

GRADE considerations for judging inconsistency (heterogeneity) in prognostic studies include variability in point estimates, extent of overlap of confidence intervals, and – for contextualized ratings - where absolute risk point estimates lie in relation to clinical decision thresholds4. There can, however, be one important difference: statistical measures of heterogeneity, such as I2, are much less useful when large studies are involved, which may be the case for observational studies addressing prognosis. Here confidence intervals are frequently narrow which may result in high I2, implying statistical heterogeneity in the absence of what would constitute important inconsistency.3

Systematic review authors should nevertheless prepare for substantial inconsistency by generating a priori hypotheses that may explain the heterogeneity they encounter. Reviewers may define substantial heterogeneity through visual inspection of individual point estimate and 95% CI of individual studies in relation to decision thresholds. In the non-contextualized setting, if reviewers are rating the certainty that the prognostic effect (i.e. risk ratio, odds ratio or hazard ratio) varies from 1.0, they will not rate down as long as all studies suggest some degree of association, whatever the magnitude. In the contextualized setting, authors may conclude substantial heterogeneity when the course of action taken through the consideration of prognosis is liable to differ between the individual studies.

When substantial heterogeneity is observed, authors should determine the extent to which their hypotheses explain the inconsistent results (appendix). Only when such exploration proves fruitless and substantial unexplained heterogeneity remains, should authors consider rating down the certainty of the evidence for inconsistency.

For instance, Witlox et al. investigated the association between delirium and post-discharge mortality in a meta-analysis of 7 observational studies, each of which reported an adjusted hazard ratio21, and reported that the presence of delirium was associated with an almost doubling of the hazard of dying (HR 1.95; 95% CI 1.51 to 2.52). In six of the eligible studies, the 95% confidence intervals varied widely. One study, however, reported a notably higher hazard ratio, with minimal overlap of 95% confidence interval with the other studies (HR 4.04; 95% CI 2.19 to 7.46).

In deciding whether to rate down in such a situation, one should consider the contribution of the aberrant study to the pooled estimate – in this case responsible for only 12% of the weight. The low weight suggests that rating down for inconsistency is unnecessary, a conclusion supported by a sensitivity analysis demonstrating similar pooled estimates whether or not one includes the aberrant study (figure 1).

**5.0 Imprecision**

Judging imprecision represents the key area in which non-contextualized and contextualized ratings are lia-ble to differ. In the former setting, authors will rate their certainty that the prognostic effect differs from a relative effect of 1.0. They may choose a different threshold – such as a relative effect of 1.5 - but finding a rationale for such an alternative threshold is likely to be challenging, and we will restrict our discussion to a threshold of 1.0.

In the contextualized setting, authors will make their decision regarding precision based on the relation between the confidence interval and a clinical decision threshold. If the clinical action would not change when the estimate is at the lower versus higher boundary of the confidence interval, the risk estimate is sufficiently precise, irrespective of the width of the interval.

Because the implications of relative estimates will differ depending on baseline risk (a RR of 2.0 may increase one’s risk from 1 to 2% or 20% to 40%) clinical decision thresholds must be expressed as absolute risks, and the most directly applicable measure of association for a prognostic factor would be a risk difference (see appendix for guidance for calculation of absolute risks). If the boundaries of the confidence interval around the absolute risk difference lie on the same side of the decisional threshold, there is no important imprecision and no reason for rating down the certainty of the evidence.

In a systematic review and meta-analysis of predictors for graft loss (defined as re-transplantation or return to dialysis) after kidney transplantation, authors identified delayed graft function as a significant prognostic factor24. The pooled hazard ratio from 4 observational studies suggested an 89% increase in risk of graft loss in patients with delayed graft function (HR 1.89, 95% CI of 1.46 to 2.47). In the non-contextualized setting, the confidence interval excludes a HR of 1.0 by a considerable margin, indicating there is no need to rate down for imprecision.

The absolute risk of 1-year graft loss in patients without delayed graft function is 60 per 1000. In those with delayed graft function, the absolute risk increases to 100 per 1000 (40 more, 95% CI of 30 more to 70 more). The contextualized setting requires asking: might clinical decisions differ between a 3% and a 7% likelihood of graft loss. Perhaps not, and if so there is no need to rate down for imprecision. The authors of the systematic review chose 5% (50 per 1000) as the risk difference threshold that might influence clinical decision-making. If one accepted that threshold, one would rate down for imprecision when considering delayed graft function.

The final decision regarding appropriate thresholds of risk for decision-making may differ for the systematic review authors and guideline panels resulting in varying judgments for imprecision. If systematic reviewers present data appropriately, it will then be possible – indeed, potentially useful - to examine the implications of different thresholds. Reporting absolute risk in people with or without the prognostic factors represents the most critical step.

**6.0 Indirectness**

Systematic review authors, guideline developers, and other evidence users need to consider whether the populations and outcomes studied correspond to their population and outcome of interest. GRADE refers to these issues as directness.

For prognostic factors, indirectness might originate when the care provided in a target population is sufficiently different from the way the patients were managed in studies included in the SR one is appraising. In the non-contextualized setting, the target population might be those entered in a clinical trial, those included in an observational study addressing intervention impact, or the potential target population for a clinical decision guide. In the contextualized setting, one would focus on the clinical target group. Similarly, if the studied outcome is not fully representative of the outcome of interest, one may consider rating down for indirectness.

In the systematic review addressing prognostic factors for graft loss at 1-year post kidney transplantation, authors meta-analyzed two cohort studies examining the association between donor’s creatinine and 1-year graft loss – which ideally would exclude mortality. The two studies, however, combined return to dialysis and re-transplantation with all-cause mortality resulting in a pooled hazard ratio of 0.95 (95% CI 0.84 to 1.07) for every 1 mg/dL increase in serum creatinine. Including mortality in the composite outcome may have contributed to the failure to demonstrate an association, a concern that warrants rating down for indirectness both for a clinical trialist designing an RCT studying an intervention to prevent graft loss, and the clinician counselling a patient regarding the likelihood of graft loss.

**7.0 Publication bias**

Publication bias may be as frequently a problem for prognostic factor research as for other research fields2 25-28. An approach that may decrase risk of overestimating associations involves searching for studies in which the prognostic factor of interest is only one of many other factors assessed. Selective reporting, however, may still bias evidence from studies reporting multiple prognostic factors2.

As is the case for overall prognosis of large groups, tests for small study effects – of which publication bias is one cause - that normalize the distribution (e.g. Begg’s test29) may be useful. Other tests, e.g. Debray’s test30, Peter’s tests31, and Egger’s test32, are applicable when heterogeneity is low and data are normally distributed.

Beyond using a statistical test, visual exploration of the funnel plot may be helpful. For example, a systematic review by Vasilevska et al. investigated the risk of cervical cancer in indigenous versus non-indigenous women33 and found a relative risk of 2.11 (95% CI 1.60 to 2.78). The authors concluded publication bias based on a positive Egger’s test for the outcome “cervical cancer not otherwise specified”. The funnel plot demonstrates, however, that the missing small studies, had they been present, would show strong positive associations (figure 2). If one believed that selective publication of small negative studies is unlikely - neither authors with small studies failing to find an association, nor editors considering a manuscript with that conclusion, are likely to be enthusiastic regarding publication - one would disregard the positive Egger’s test and not rate down for publication bias.

**8.0 Rating up certainty**

Although we have thus far not found examples of systematic reviews that mandate rating up for prognostic factors based on dose response, large effect or for the nature of plausible biases, examples may emerge with further experience. For instance, one may imagine a non-contextualized scenario in which we observe a very strong association (>5 or <0.2) with no concerns of risk of bias or imprecision. Under such circumstances, it may be warranted to increase the certainty rating.

**9.0 Additional/Cautionary remarks**

Special considerations arise when simultaneously considering more than one GRADE domain. For instance, one potential hypothesis for exploration of inconsistency can be risk of bias. It may be that authors observe no significant difference in the effect estimate observed from studies classified as high and low risk of bias. Therefore, authors may opt to maintain high risk of bias studies within their analysis. With regards to inconsistency, however, authors may decrease their certainty in the evidence due to large inconsistency amongst high risk of bias studies. In such a case, it may be reasonable for authors to exclude high risk of bias studies, not because of biased associations, but rather because of inconsistent associations. The review by Cheng et al. provides an example20.

In the review addressing the association between smoking and VTE20, investigators observed substantial differences in point estimates of relative risk from 0.50 to 4.70 with limited overlap in confidence intervals from individual studies (figure 3). They postulated that the association between smoking and VTE would differ when the primary studies were adjusted for other prognostic factors. Indeed, a sensitivity analysis including only adjusted relative risks resulted a more visually consistent forest plot with the individual study associations ranging from RR 0.90 to 2.87 (figure 3).

The pooled effect estimates from unadjusted and adjusted studies did not, however, differ significantly (unadjusted RR: 1.17, 95% CI 1.09 to 1.25; adjusted RR: 1.21, 95% CI 1.15 to 1.26). Therefore, it is reasonable that the authors maintained the unadjusted studies in the primary analysis. With regards to inconsistency, however, a separate meta-analysis for adjusted studies does considerably reduce heterogeneity and thus warrants using only adjusted studies without rating down for inconsistency.

Our overall judgment on certainty for the effect of a particular factor may differ for non-contextualised and contextualised settings. In particular, it may be possible that risk of bias may be low for pooled relative effects (i.e. non-contextualised settings), but high for pooled absolute risk predictions for groups defined by a prognostic factor (i.e. contextualised settings). The reason is that estimates of absolute risk predictions are more prone to overfitting concerns; that is, the estimated predicted probabilities are too close to 0 or 1. This issue is noted in PROBAST13. The magnitude and frequency of this risk of bias is uncertain, and currently represents a limitation of our approach to rating the certainty of the evidence in the contextualized framework. Ideally, absolute risk predictions conditional on a prognostic factor would address overfitting by using penalisation and shrinkage techniques, and examine calibration in new data. This, however, is rarely done. We plan to further consider this issue in subsequent work on GRADE for prognostic model research.

**10.0 Concluding remarks**

The same principles GRADE proposed for bodies of evidence addressing treatment and overall prognosis work well in assessing individual prognostic factors, both in non-contextualized and contextualized settings. Using GRADE guidance, and documenting the logic of its application, will ensure clinical investigators and clinicians understand the certainty associated with the evidence, and the rationale for certainty ratings:.

**Competing interest statement**

All authors declare they did not receive support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years exist, nor do other relationships or activities that could appear to have influenced the submitted work. All authors are members of the GRADE working group.

**Contributions**

All authors contributed to the generation of the research hypothesis, participated to the discussion of its content and approved the final version of the manuscript. FF, VZ, and AI selected the systematic reviews used as examples and prepared summary of finding tables used in the process. FF drafted the manuscript, AI is the guarantor.

**Ethical approval**

Not applicable.

**Funding**

No external funding

**Data sharing**

No additional data available.

**References**

1. Riley RD, van der Windt DA, Croft P, et al. Prognosis Research in Health Care: Concepts, Methods, and Impact: Oxford University Press 2019.

2. Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10(2):e1001380. doi: 10.1371/journal.pmed.1001380 [published Online First: 2013/02/09]

3. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ* 2015;350:h870. doi: 10.1136/bmj.h870 [published Online First: 2015/03/18]

4. 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. doi: 10.1016/j.jclinepi.2017.05.006 [published Online First: 2017/05/23]

5. Altman DG. Systematic reviews of evaluations of prognostic variables. *BMJ* 2001;323(7306):224-8. [published Online First: 2001/07/28]

6. Hemingway H, Philipson P, Chen R, et al. Evaluating the Quality of Research into a Single Prognostic Biomarker: A Systematic Review and Meta-analysis of 83 Studies of C-Reactive Protein in Stable Coronary Artery Disease. *PLOS Medicine* 2010;7(6):e1000286. doi: 10.1371/journal.pmed.1000286

7. Matino D, Chai-Adisaksopha C, Iorio A. Systematic reviews of prognosis studies: a critical appraisal of five core clinical journals. *Diagnostic and Prognostic Research* 2017;1(1):9. doi: 10.1186/s41512-017-0008-z

8. Riley RD, Moons KGM, Snell KIE, et al. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ* 2019;364:k4597. doi: 10.1136/bmj.k4597 [published Online First: 2019/02/01]

9. Brożek JL, Akl EA, Jaeschke R, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy* 2009;64(8):1109-16. doi: doi:10.1111/j.1398-9995.2009.02083.x

10. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336(7653):1106-10. doi: 10.1136/bmj.39500.677199.AE [published Online First: 2008/05/17]

11. Hordijk-Trion M, on behalf of the EHSCRI, Lenzen M, et al. Patients enrolled in coronary intervention trials are not representative of patients in clinical practice: results from the Euro Heart Survey on Coronary Revascularization. *European Heart Journal* 2006;27(6):671-78. doi: 10.1093/eurheartj/ehi731

12. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158(4):280-6. doi: 10.7326/0003-4819-158-4-201302190-00009 [published Online First: 2013/02/20]

13. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170(1):51-58. doi: 10.7326/M18-1376 [published Online First: 2019/01/01]

14. Reeves BC, Deeks JJ, Higgins JP, et al. Including non-randomized studies. Higgins JP, Green S, editors Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration 2010:391–432.

15. Cook DJ, Fuller HD, Guyatt GH, et al. Risk Factors for Gastrointestinal Bleeding in Critically Ill Patients. *New England Journal of Medicine* 1994;330(6):377-81. doi: 10.1056/nejm199402103300601

16. Kucher N, Kohler HP, Dornhofer T, et al. Accuracy of D-dimer/fibrinogen ratio to predict pulmonary embolism: a prospective diagnostic study. *J Thromb Haemost* 2003;1(4):708-13. [published Online First: 2003/07/23]

17. Perrier A, Roy PM, Sanchez O, et al. Multidetector-row computed tomography in suspected pulmonary embolism. *N Engl J Med* 2005;352(17):1760-8. doi: 10.1056/NEJMoa042905 [published Online First: 2005/04/29]

18. Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997;350(9094):1795-8. doi: 10.1016/S0140-6736(97)08140-3 [published Online First: 1998/01/15]

19. Sanchez O, Trinquart L, Colombet I, et al. Prognostic value of right ventricular dysfunction in patients with haemodynamically stable pulmonary embolism: a systematic review. *Eur Heart J* 2008;29(12):1569-77. doi: 10.1093/eurheartj/ehn208 [published Online First: 2008/05/23]

20. Cheng YJ, Liu ZH, Yao FJ, et al. Current and former smoking and risk for venous thromboembolism: a systematic review and meta-analysis. *PLoS Med* 2013;10(9):e1001515. doi: 10.1371/journal.pmed.1001515 [published Online First: 2013/09/27]

21. Witlox J, Eurelings LS, de Jonghe JF, et al. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA* 2010;304(4):443-51. doi: 10.1001/jama.2010.1013 [published Online First: 2010/07/29]

22. Schandelmaier S, Briel M, Varadhan R, et al. Development of a new Instrument to assess the Credibility of Effect Modification ANalyses (ICEMAN) in randomized controlled trials and meta- analyses. McMaster University, 2019.

23. Foroutan F, Guyatt G, Thabane L, et al. Calculation of absolute risk for important outcomes in patients with and without a prognostic factor of interest. *Journal of Clinical Epidemiology* 2019;In press

24. Foroutan F, Friesen EL, Clark KE, et al. Risk Factors for One-Year Graft Loss Post-Kidney Transplantation. *Clinical Journal of the American Society of Nephrology* 2019;In Press

25. Andre F, McShane LM, Michiels S, et al. Biomarker studies: a call for a comprehensive biomarker study registry. *Nature Reviews Clinical Oncology* 2011;8:171. doi: 10.1038/nrclinonc.2011.4

26. Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer* 2007;43(17):2559-79. doi: 10.1016/j.ejca.2007.08.030

27. Martin B, Paesmans M, Berghmans T, et al. Role of Bcl-2 as a prognostic factor for survival in lung cancer: a systematic review of the literature with meta-analysis. *British Journal Of Cancer* 2003;89:55. doi: 10.1038/sj.bjc.6601095

28. McShane LM, Altman DG, Sauerbrei W. Identification of Clinically Useful Cancer Prognostic Factors: What Are We Missing? *JNCI: Journal of the National Cancer Institute* 2005;97(14):1023-25. doi: 10.1093/jnci/dji193

29. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;50(4):1088-101. [published Online First: 1994/12/01]

30. Debray TPA, Moons KGM, Riley RD. Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests. *Res Synth Methods* 2018;9(1):41-50. doi: 10.1002/jrsm.1266 [published Online First: 2017/10/05]

31. Peters JL, Sutton AJ, Jones DR, et al. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006;295(6):676-80. doi: 10.1001/jama.295.6.676 [published Online First: 2006/02/10]

32. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34. [published Online First: 1997/10/06]

33. Vasilevska M, Ross SA, Gesink D, et al. Relative risk of cervical cancer in indigenous women in Australia, Canada, New Zealand, and the United States: a systematic review and meta-analysis. *J Public Health Policy* 2012;33(2):148-64. doi: 10.1057/jphp.2012.8 [published Online First: 2012/03/02]

34. Thompson M, Vodicka TA, Blair PS, et al. Duration of symptoms of respiratory tract infections in children: systematic review. *BMJ* 2013;347:f7027. doi: 10.1136/bmj.f7027 [published Online First: 2013/12/18]

**Table 1. Types and goals of prognostic studies**

|  |  |  |
| --- | --- | --- |
| **Study type** | **Study Goal** | **Examples in the field of atrial fibrillation** |
| Overall prognosis(#) | Establish the typical risk in a broadly defined population | Risk of bleeding in patients with atrial fibrillation receiving vitamin K antagonists |
| Prognostic factor | Establish how a particular patient characteristic influences risk | Influence of age on the risk of bleeding in patients with atrial fibrillation |
| Outcome (or risk) prediction model | Development of a full prognostic model simultaneously considering a number of prognostic factors and classifying patients into various levels of risk | CHADS2 and CHADS-VASC for the risk of strokeHAS-BLED, HEMORRHAGE for the risk of bleeding |

(#) – It is equally important to estimate the likelihood of spontaneous resolution of a disease, as discussed in Matthew Thompson et al34.

**Table 2 – Significance of levels of evidence for risk associated with the prognostic factor**

|  |  |
| --- | --- |
| Quality level | Definition |
| High  | We are very confident that the variation in risk associated with the prognostic factor (probability of future events in those with/without the prognostic factor) lies close to that of the estimate (#) |
| Moderate | We are moderately confident that the variation in risk associated with the prognostic factor (probability of future events in those with/without the prognostic factor) is likely to be close to the estimate, but there is a possibility that it is substantially different |
| Low | Our certainty in the estimate is limited: The variation in risk associated with the prognostic factor (probability of future events in those with/without the prognostic factor) may be substantially different from the estimate |
| Very low | We have very little certainty in the estimate: The variation in risk associated with the prognostic factor (probability of future events in those with/without the prognostic factor) is likely to be substantially different from the estimate |

# Prognostic factor studies measure the variation in incidence, i.e. target events over time in a population of interest at risk for the target event, as a function of presenting or not a specific prognostic factor. The target event can be an adverse outcome (e.g. mortality) in patients with a prognostic factor as compared to those without (e.g. BMI > 30 as compared to <25). Sometimes the prognostic factor is the change in some patient characteristic over time (e.g. the combination of the Apgar score at 1 and 5 minutes after delivery).

Figure 1 – Primary (left) and sensitivity (right) analysis for the association between delirium and mortality

Figure 2 – Funnel plot for visual inspection of publication bias in a review on risk cervical cancer in indigenous versus non-indigenous women.

Figure 3 – Primary and subgroup analyses for association between smoking and venous thromboembolism. Subgroup based on adjustment for confounders.