ORIGINAL ARTICLE

GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence

Holger J. Schünemann[a,b,*], Carlos Cuello[a], Elie A. Akl[a,c], Reem A. Mustafa[a,d],
Jörg J. Meerpohl[e], Kris Thayer[f], Rebecca L. Morgan[a], Gerald Gartlehner[g], Regina Kunz[h],
S Vittal Katikireddi[i], Jonathan Sterne[j], Julian PT Higgins[j], Gordon Guyatt[a,b],
GRADE Working Group

[a]Department of Health Research Methods, Evidence, and Impact and McGRADE Center, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S4K1, Canada
[b]Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S4K1, Canada
[c]AUB GRADE Center, Clinical Research Institute, American University of Beirut, PO Box 11-0236, Riad El Solh, Beirut, 1107 2020, Lebanon
[d]Department of Medicine, University of Kansas Medical Center, 3901 Rainbow Blvd, MS3002, Kansas City, KS, 66160, USA
[e]Cochrane Germany, Medical Center University of Freiburg, Breisacher Strasse 153, Freiburg, 79110, Germany
[f]Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment, Environmental Protection Agency, 1200 Pennsylvania Avenue, N.W.Washington, DC 20460, USA
[g]Department for Evidence-Based Medicine and Clinical Epidemiology, Danube University Krems, Dr Karl Dorrek Straße 30, Krems, 3500, Austria
[h]Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, Basel, 4031, Switzerland
[i]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Top Floor, 200 Renfield Street, Glasgow, G2 3QB, Scotland
[j]Population Health Sciences, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK

Abstract

Objective: To provide guidance on how systematic review authors, guideline developers, and health technology assessment practitioners should approach the use of the risk of bias in nonrandomized studies of interventions (ROBINS-I) tool as a part of GRADE's certainty rating process.

Study Design and Setting: The study design and setting comprised iterative discussions, testing in systematic reviews, and presentation at GRADE working group meetings with feedback from the GRADE working group.

Results: We describe where to start the initial assessment of a body of evidence with the use of ROBINS-I and where one would anticipate the final rating would end up. The GRADE accounted for issues that mitigate concerns about confounding and selection bias by introducing the upgrading domains: large effects, dose-effect relations, and when plausible residual confounders or other biases increase certainty. They will need to be considered in an assessment of a body of evidence when using ROBINS-I.

Conclusions: The use of ROBINS-I in GRADE assessments may allow for a better comparison of evidence from randomized controlled trials (RCTs) and nonrandomized studies (NRSs) because they are placed on a common metric for risk of bias. Challenges remain, including appropriate presentation of evidence from RCTs and NRSs for decision-making and how to optimally integrate RCTs and NRSs in an evidence assessment. © 2018 Elsevier Inc. All rights reserved.

Keywords: GRADE; Quality of evidence; Certainty of the evidence; Risk of bias; Nonrandomized studies; ROBINS

1. GRADE's approach to rate the certainty of the evidence from observational studies

The GRADE working group has developed a widely accepted approach to rate the certainty of a body of evidence (also known as quality of evidence or confidence in evidence) in the contexts of systematic reviews, developing health-care recommendations, and supporting decisions.

What is New?

Key findings

- The risk of bias in nonrandomized studies of interventions (ROBINS-I) tool addresses risk of bias in relation to a randomized trial presenting a number of opportunities for the GRADE approach.

- The GRADE working group addresses here how tools like ROBINS-I should be used to assess risk of bias in observational or nonrandomized studies (NRS).

- The GRADE approach already accounted for issues that mitigate concerns about confounding and selection bias by introducing the upgrading domains: large effects, dose-effect relations, and when plausible residual confounders or other biases increase certainty.

What this adds to what is known?

- In GRADE, the separation of randomized and observational studies was primarily a result of recognition that randomization is the only way to fully protect against confounding and that confounding is always a concern in even the most rigorously conducted observational studies.

- Using the term NRSs clarifies that all other designs but randomized trials are mentioned when evaluating observational studies in GRADE.

What are the implications and what should change now?

- The use of ROBINS-I in GRADE assessments may allow for a better comparison of evidence from randomized controlled trials and NRSs because they are placed on a common metric for risk of bias.

- The article describes the initial assessment of a body of evidence with the use of ROBINS-I and where one would anticipate the final rating would end up.

GRADE's approach to rating the certainty of the evidence is based on a four-level system: high, moderate, low, and very low (Table 1). This is the 18th in the ongoing series of articles describing the GRADE approach in the Journal of Clinical Epidemiology and complements articles in other journals. In previous GRADE articles, we have described the reasons for decreasing and increasing the certainty of a body of evidence; how an overall rating of the evidence is performed; how evidence is used to move to recommendations and decisions; how to deal with particular

circumstances of diagnostic, prognostic, equity-related, multiple treatment comparison, environmental, and public health questions; how GRADE applies to rapid advice; and when there is missing outcome data [1–17].

The current GRADE approach for a body of evidence relating to interventions begins by placing studies in one of two categories: randomized controlled trials (RCTs) and observational studies (otherwise known as nonrandomized studies [NRSs]). GRADE considers nonrandomized trials, cohort studies, case-control studies, interrupted time series (if not randomized), cross-sectional studies, case series, case reports, and other types of NRSs as observational studies.

According to existing GRADE guidance for interventions, the process of rating a body of evidence (typically several or many studies) begins by classifying the design used by the relevant studies. If the relevant studies are randomized trials, the body of evidence begins as high certainty. If the relevant studies are observational, the body of evidence begins as low certainty. This initial rating is followed by consideration of eight domains, five of which may result in rating down certainty and three in rating up [8].

The separation of randomized and observational studies was primarily a result of recognition that randomization is the only way to fully protect against confounding (i.e., imbalance in prognostic factors between intervention and control groups) and that confounding is always a concern in even the most rigorously conducted observational studies. The imbalance in unknown prognostic factors that exists after statistical adjustment or stratified analysis to account for known variables which are not balanced in the exposed and control groups is known as residual confounding.

The choice of starting observational studies at low rather than moderate or very low certainty followed intense discussion in the GRADE working group's early days and was based on the group's assessment of the magnitude of the potential for residual confounding and the limited protection against bias provided by adjusted analysis in observational studies. An alternative way of understanding GRADE is that randomization is one of the reasons for rating up certainty as a measure to protect against confounding and selection bias.

2. Rating risk of bias in individual observational studies

Consider now the assessment of risk of bias in individual observational studies, which in the GRADE approach might lead to further rating down quality from low to very low. Investigators have developed many assessment tools for rating risk of bias in observational studies. Most of the instruments address a specific type of observational or nonrandomized design (e.g., cohort or case control) [18] and seek to determine how well, relative to a perfect observational study of that particular design, the individual study at hand was conducted. An alternative approach is to

Table 1. Use of GRADE not considering ROBINS-I and similar tools: According to GRADE, certainty, quality, strength of the evidence, or the confidence in the estimate of effect, is determined for each outcome based on a systematic review of the evidence for each outcome. For recommendations, the overall certainty is determined across outcomes based on the lowest quality outcome among those critical for decision-making for the specific context.

1. Establish initial level of certainty (as implemented in current GRADE)		2. Consider lowering or raising level of certainty		3. Final level of certainty rating
Study design	*Initial certainty in the evidence*	*Reasons for considering lowering or raising certainty*		*Certainty in the evidence across those considerations*
		⬇ Lower if	⬆ Higher if*	
Randomized trials ➜	High certainty	Risk of Bias	Large effect	High ⊕⊕⊕⊕
		Inconsistency	Dose response	
		Indirectness	All plausible confounding and bias	Moderate ⊕⊕⊕◯
		Imprecision	• would reduce a demonstrated effect **or**	
Observational studies ➜	Low certainty	Publication bias	• would suggest a spurious effect if no effect was observed	Low ⊕⊕◯◯
				Very low ⊕◯◯◯

*Criteria for upgrading the quality are usually only applicable to observational studies without any reason for rating down.

determine risk of bias of observational studies in relation to the effect that would be seen in a high-quality randomized trial. Such a trial avoids both confounding (through random allocation to interventions) and other sources of bias such as selection or information biases.

The risk of bias in nonrandomized studies of interventions (ROBINS-I) tool, rather than using the ideal observational study as a standard, addresses risk of bias using an absolute scale approach [19]. The ROBINS-I evaluates risk of bias in estimates of the effects (harm or benefit) of one or more interventions from studies that did not use randomization to allocate units (individuals or clusters of individuals) to compare groups (in GRADE terminology observational studies).

ROBINS-I's fundamental underlying principles are that [1] a study's risk of bias is compared against a target RCT, even if this RCT may not be feasible or ethical [2]; the assessment of confounding and selection bias are integral parts of the tool; and [3] for a given result, for a specific outcome, evidence from an NRS is assessed, addressing a number of domains and then giving an overall rating per outcome for each study. Fig. 1 describes the application of ROBINS-I. Signaling questions in the ROBINS-I instrument ask respondents to rate risk of bias in domains of (1) bias due to confounding, (2) bias in selection of participants into the study, (3) bias in classification of interventions, (4) bias due to departures from intended interventions, (5) bias due to missing data, (6) bias in measurement of outcomes, and
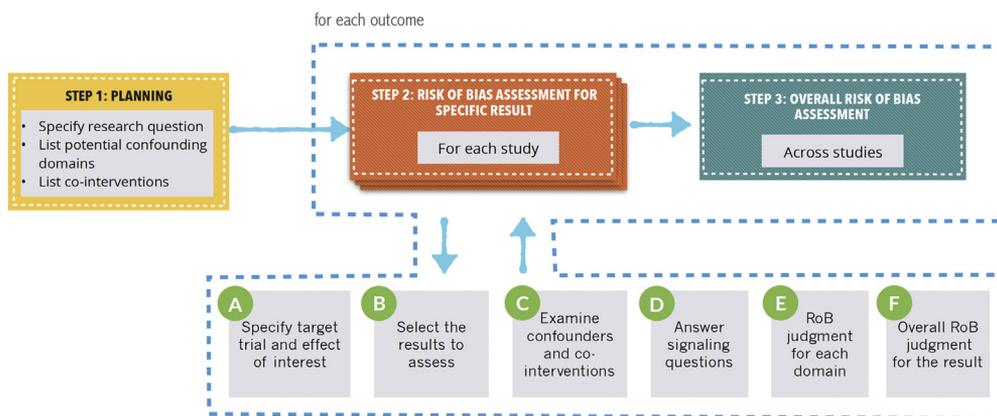

Fig. 1. The process for using ROBINS-I. ROBINS-I, risk of bias in nonrandomized studies of interventions.
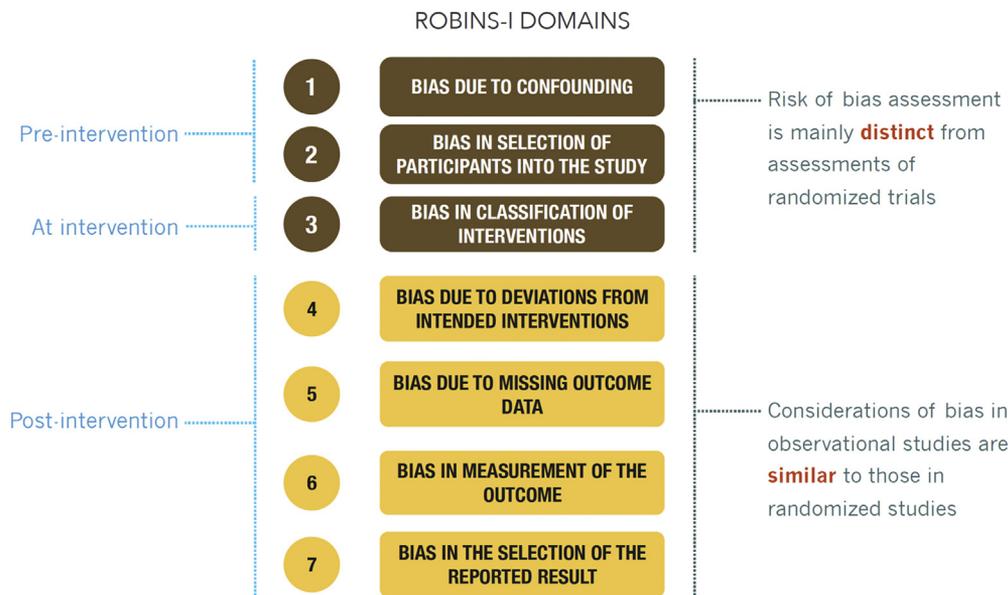
ROBINS-I DOMAINS

Fig. 2. ROBINS-I risk of bias domains. In GRADE, risk of bias is a domain, and ROBINS-I domains are called items. ROBINS-I, risk of bias in nonrandomized studies of interventions.

(7) bias in selection of reported results (Fig. 2). In addition, ROBINS-I includes an optional judgment about the direction of the bias for each domain. ROBINS-I has undergone careful development by a large group of experienced investigators. It has been tested, and scientists have begun to validate it; therefore, experience will continue to accumulate. Box 1 provides a clarification about terminology used in the ROBINS tool and by GRADE.

3. ROBINS-I and GRADE

The arrival of ROBINS-I presents a number of opportunities for the GRADE approach. First, it offers an alternative terminology: establishing NRS rather than observational studies. Although not different in intended meaning in the GRADE approach, substituting NRS for observational studies will lead to a more transparent separation of studies based on their design. For instance, some have struggled with the classification of certain types of studies, such as nonrandomized before-after studies, as observational; in the alternative nomenclature, such studies are clearly nonrandomized. How to classify studies that are allocated by essentially random processes such as date of birth or hospital ID number, in which the concern is lack of concealment rather than confounding bias per se, may remain a matter of debate that we will not address here. We will use the term NRS from here on in this article.

Second, the use of ROBINS-I in GRADE assessments may allow for a better comparison of evidence from RCTs and NRS because they are placed on a common metric for risk of bias. This article provides guidance regarding how systematic review authors, guideline developers, and health technology assessment practitioners using GRADE might approach the use of ROBINS-I as part of the certainty

rating process. The article focuses on where to start the initial assessment of a body of evidence with the use of ROBINS-I and where one would anticipate the final rating would end up. Implications and requirement for further work are dealt with in the final sections of this article. This article will not resolve all relevant issues, and we have planned for subsequent articles describing the work of the GRADE RoB in NRS and environmental health project groups (www.gradeworkinggroup.org).

4. Concerns about GRADE's approach to start an NRS at low certainty

Despite GRADE's broad acceptance in the evidence synthesis community, GRADE's initial certainty rating of outcome data from NRS as low has led to challenges for some GRADE users. First, users of GRADE may inappropriately double count the risk of confounding and selection bias, initially by starting a body of evidence from NRS as low certainty of the evidence followed by again rating down for unknown confounders (although rating down additionally for failure to accurately measure known confounders and adjusting for these confounders in the analysis would be appropriate [Fig. 3]). Second, those working in fields in which RCTs are sparse or not feasible have expressed concerns that NRS in their fields will seldom be rated as high or perhaps even moderate certainty. GRADE has accepted that criticism and highlighted how one may rate up certainty for large effects, a dose-response gradient, if all plausible biases will strengthen rather than undermine inferences from study results. In this article, we noted the merits of a rating system that follows the underlying logic of ROBINS-I and thus may better integrate RCTs and NRS and allow for more detailed assessment of different types of NRS.

Determinants of certainty of evidence

- RCTs ⊕⊕⊕⊕ | **high**
- **observational studies** ⊕⊕◯◯ | **low**		Risk of bias
- **5 domains that can lower certainty**
  1. limitations in detailed study design and execution (*risk of bias* items)
  2. Inconsistency *(or heterogeneity)*
  3. Indirectness *(PICO and applicability)*
  4. Imprecision
  5. Publication bias
- **3 domains can increase certainty**
  1. large magnitude of effect
  2. opposing plausible residual bias or confounding
  3. dose-response gradient

Fig. 3. The current GRADE approach for certainty of evidence: initial certainty and rating domains.

Although best evidence must be used for decision-making, relying on the best available or achievable rather than least biased evidence as a reference standard would lead to differing certainty in decisions based on the questions asked [20]. Picture the following: in one scenario for a health-care decision, RCTs are neither ethical nor feasible, and we, therefore, accept that possible confounded NRSs are the reference standard for highest feasible certainty. If these studies are available, we would express that we have high certainty in the decision despite the fact that confounding may bias the results. In the second scenario, RCTs are feasible and ethical, and they become our reference standard for highest feasible certainty. If for this situation, only NRSs are available, we would label our health-care decision as based on low certainty. Should the certainty of the decision that is based on the respective evidence differ because of what evidence is available or should the certainty depend on what would be the highest possible certainty? It would be illogical to express a different certainty for the same degree of bias because of feasibility and ethical reasons. A comparison on an absolute rather than relative scale (to the feasibility and ethics of an RCT) provides greater transparency. A decision can still be made for both the scenarios, and for both, we should acknowledge the (same) degree of uncertainty. In fact, in most, if not all, areas of health care, some interventions are supported by evidence from RCTs and others are not, requiring a common reference standard to ensure appropriate communication with target populations.

Third, by beginning the rating of evidence from a body of NRS as low certainty, the current GRADE approach fails to consider that a body of evidence from particular NRS designs may more appropriately be rated higher than conventional NRS designs. For instance, interrupted time series with multiple periods and measurements during each period with no other limitations may constitute moderate-quality evidence without meeting any of the criteria for rating up (though our efforts to identify examples for such a body of evidence have not yet proved successful) [21].

5. Certainty of evidence for a body of evidence from NRS when using ROBINS-I for assessing risk of bias in individual NRSs

Here, we provide general guidance for the use of GRADE in the context of ROBINS-I. ROBINS-I compares an assessment of an individual NRS against a target RCT. The initial description of the underlying study design, such as cohort, case-control, case series, or cross-sectional study, is not considered as a risk of bias feature in ROBINS-I. Thus, when using ROBINS-I for assessing risk of bias in NRS, given that assessment of selection bias and confounding is an integral part of the ROBINS-I tool, the initial GRADE certainty in the evidence from a body of studies using an NRS design would be high (Figs. 4 and 5). This does not mean that GRADE has changed the view that randomization is the only secure way to guard against confounding bias; that view remains the same. Thus, we would anticipate whether one begins with a body of evidence from NRSs as low certainty and looks for reasons to rate up or down or starts with that evidence as high quality and looks for reasons to rate down, the final certainty rating should be the same.

This approach implies that ROBINS-I users rating conventional NRS of any design (e.g., cohort, case-control) after their assessment of confounding and selection bias, will often arrive at a rating of high risk of bias. Using ROBINS-I, it nevertheless remains possible that a body of evidence from NRSs will receive a final rating of high or moderate certainty of evidence. This could result from rating up for large effect, dose response, or the direction of plausible confounding. Or, it could result from the use of NRS designs and analyses with greater protection against risk of bias—for instance, interrupted time series—that would lead to rating down by only one level or not at all. As we have already noted, however, that although we have many examples of rating up certainty, efforts to identify a body of evidence from innovative designs meriting, simply for design considerations, moderate-quality evidence, have thus far proved unsuccessful (Fig. 4). Methodological developments in this area, which describe how NRS may have greater protection against risk of bias than those typically available, should help making such judgments, but GRADE requires careful examination of these examples [22].

6. What makes us confident in results of NRS and does GRADE already account for this?

At the end of the previous section, we have noted how, within current GRADE thinking, a body of evidence from NRSs may emerge from the rating exercise as moderate or high-quality evidence. We will now discuss on these issues.

6.1. All plausible residual confounders or other biases increase our certainty in the estimated effect

GRADE allows higher certainty ratings for bodies of evidence when all plausible residual confounders or other
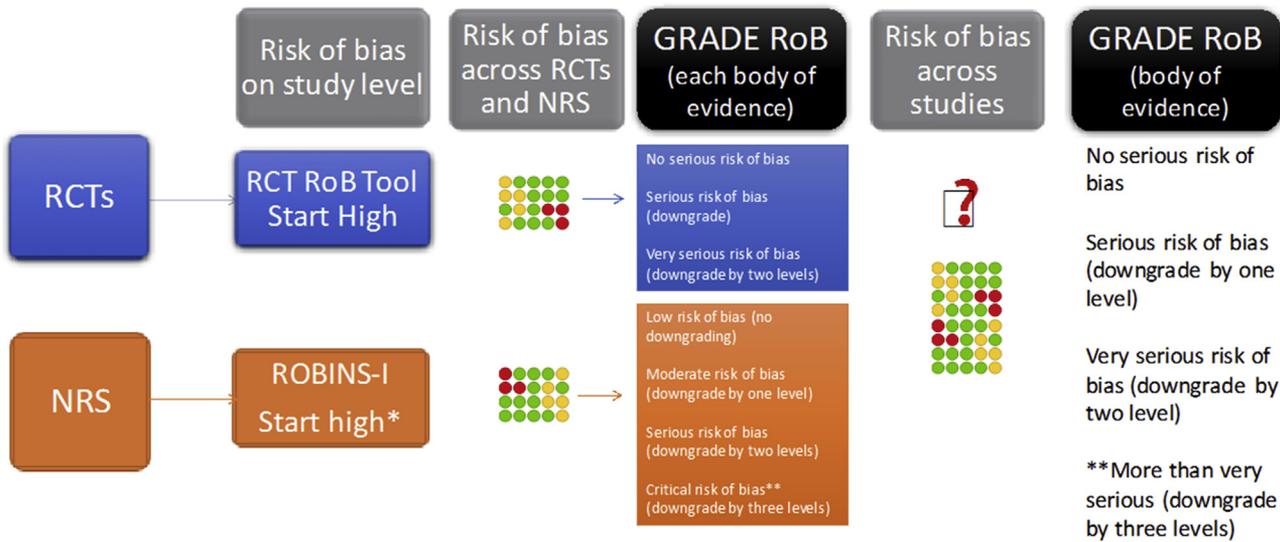
Fig. 4. Assessing randomized trials and nonrandomized studies with GRADE. Rating of risk of bias for randomized and nonrandomized studies. Different tools are used to address for the two types of bodies of evidence (randomized and nonrandomized). The risk of bias is then assessed across studies and integrated if possible. When ROBINS-I is used, rating down by more than two levels is required for a body of evidence that is rated as critical on that instrument. *In practice, this will generally lead to rating down by at least two levels to low or very low certainty for NRSs. However, for results with large effects or dose response or results in which inference is strengthened by the plausible biases that exist, the extent of rating down may be lowered. We have not identified bodies of evidence in which a ROBINS-I assessment alone leads to no rating down or rating down by only one level. How to integrate RCTs and NRSs will be further discussed in upcoming GRADE guidance articles. **For GRADE, the corresponding terminology is not serious, serious, very serious, and a fourth level of risk of bias. GRADE is currently exploring the appropriate term for the fourth level. ROBINS-I, risk of bias in nonrandomized studies of interventions; NRSs, nonrandomized studies; RCTs, randomized controlled trials.

biases increase our confidence in the range of an estimated effect, that is, the effect is either larger or smaller than that observed [23,24]. GRADE suggests that judgments about the direction of the possible bias are important to assess certainty of the evidence from NRSs. One example from the public health field comes from a systematic review of NRSs including a total of 38 million patients who demonstrated a very small relative increase (relative risk [RR], 1.020; 95% confidence interval [CI], 1.003–1.038) in death rates in private for-profit hospitals compared with private

Certainty of evidence with tools like ROBINS-I

- **RCTs and NRS ⊕⊕⊕⊕ | high**
- **Domains that can lower certainty**
  1. limitations in detailed study design and execution *(risk of bias items)*
     - *Lack of randomization lowers certainty to low unless opposing plausible residual bias strengthens certainty or special study designs that reduce confounding and selection bias*
  2. Inconsistency *(or heterogeneity)*
  3. Indirectness *(PICO and applicability)*
  4. Imprecision
  5. Publication bias
- **Domains can increase certainty or mitigate risk of bias**
  1. large magnitude of effect
  2. dose-response gradient

Fig. 5. GRADE approach for certainty of evidence with tools like ROBINS-I. ROBINS-I, risk of bias in nonrandomized studies of interventions.

not-for-profit hospitals [23,25]. The evaluation of risk of bias across studies revealed that all residual plausible confounding—the major issues being that for-profit hospitals have on average higher income patients and greater resources—would have decreased the observed effect (further toward an RR of 1.0). Despite the biases in favor of for-profit hospitals, those hospitals demonstrated higher mortality; therefore, the true effect, if it differs from the estimate, is almost certain to be greater.

Currently, ROBINS-I allows for an optional judgment regarding the direction of confounding and selection bias (''Risk of bias judgment. Optional: What is the predicted direction of the bias due to confounding/selection …''). If this optional judgment is indeed used in ROBINS-I, in scenarios such as the hospital profit status example, not rating an individual NRS as high risk of bias (thus, not rating the body of evidence from a number of such studies as low certainty) is justified. This may happen, albeit rarely, even in the context of small effects such as the one observed for the mortality risk in for-profit private hospitals.

While GRADE has accounted for this situation in its approach, when users of GRADE apply ROBINS-I to assess risk of bias, the direction and degree of residual plausible confounding require consideration during the risk of bias assessment. Rather than rating up NRS from low to moderate at the study and body of evidence level, raters using ROBINS-I may not rate risk of bias as very serious, but only rate it as moderate. Whether or not one starts at low certainty in the traditional GRADE approach and rates

up or does not rate down to low when using ROBINS-I, the end result is identical and depends on the risk of bias judgment (Fig. 5).

6.2. Large effects and dose responses

GRADE suggests that large effects and dose-response relations mitigate concerns regarding residual confounding. In one of our prior articles, we described that a systematic review of NRSs investigating the effect of cyclooxygenase-2 inhibitors on cardiovascular events found the summary estimate of RR with rofecoxib of 1.33 (95% CI: 1.00−1.79) with doses less than 25 mg/d and 2.19 (95% CI: 1.64−2.91) with doses more than 25 mg/d. Can we infer that rofecoxib will increase the risk for cardiovascular events? Although only NRSs are available to address the question, we can have moderate, or perhaps even high, certainty of the causal connection. The reasons are that although residual confounding is likely to exist in the NRSs that address this issue, the existence of a dose-response gradient and the large apparent effect of higher doses of rofecoxib markedly increase our strength of inference that the association cannot be explained by residual confounding and is therefore likely to be both causal and, at high levels of exposure, substantial.[1] Given the large effect and the observed dose-response relation, this could lead to a high certainty rating, for the outcome of increasing cardiovascular events.

The previous paragraph dealt with evaluation of the entire body of evidence, which begs the question of how the rating of individual NRSs of rofecoxib using ROBINS-I would impact on a GRADE assessment of the body of evidence. The rater, in dealing with the confounding and selection bias domains, would rate individual study as high risk of bias because of the possibility that residual confounding or selection bias may be influencing the estimates of association. How would one then deal with the dose-response and large effect size considerations when dealing with the body of evidence?

In one way of looking at the situation, the subsequent rating up for large effects for the higher doses of rofecoxib would make, in retrospect, some of the items on the ROBINS-I tool potentially irrelevant. The possible solutions are to (1) rate the confounding in ROBINS-I as moderate or low risk of bias because large effects are observed and the larger the effect, the stronger the confounding would have to be to explain the effect which makes an explanation by confounding unlikely or (2) leave the initial grading as low following the guidance mentioned previously and then rate up for large effects when one considers

the entire body of evidence. The same options exist with respect to dose-response relationships. GRADE has thus accounted for issues that mitigate concerns about confounding and selection bias by introducing the upgrading domains. They will hence need to be considered in an assessment of a body of evidence when using ROBINS-I.

7. Advantages and disadvantages of, in the context of GRADE, assessing risk of bias for individual studies using the ROBINS-I approach of specifying target trials

7.1. Advantages

Among other features, ROBINS-I allows review authors to assess how failure to use randomization in individual studies has impacted on risk of bias. For example, ROBINS-I allows categorization of the magnitude of bias from lack of randomization through the selection and confounding bias domains, application of this assessment across risk of bias domains, and evaluation of how this differs across individual studies that address different health-care questions. Furthermore, ROBINS-I will facilitate assessment of a study that has been described as randomized, but when assessed in detail, it is found to be not appropriately randomized. In those cases, users of GRADE have struggled with whether to start the certainty of evidence as high and then rate it down or ignore descriptions of the study authors and treat the studies as NRS by starting the certainty of evidence as low. All these features of the assessment of individual studies can then be taken into account when evaluating a group of individual studies that constitute a body of evidence.

Another potential advantage of using an approach such as ROBINS-I is that it may harmonize GRADE approaches across different study types for different types of questions such as prognosis or test accuracy. In the current GRADE approach, NRSs for these types of questions begin with high certainty ratings. In particular, with prognostic studies, in which the issue is association and not causation, prognostic NRSs begin as high certainty evidence. If GRADE assessments for all types of studies were to start at high certainty, questions of intervention, prognosis, values and preferences, and test accuracy would not require different initial certainty ratings. What will be required, however, are different versions of ROBINS, such as ROBINS tool for prognosis.

Finally, those applying GRADE in fields where RCTs are sparse, such as environmental and certain areas of public health, reframing the certainty assessment with a focus on the actual items that randomization addresses, that is, confounding and selection bias, rather than labeling a study design feature, that is, randomization, will find GRADE more acceptable.

7.2. Disadvantages

Disadvantages of offering an alternative to the existing GRADE system include mistakes if users of GRADE do

---

[1] GRADE guidance suggests the possibility of rating up one level for a large effect if the relative effect is greater than 2.0. Here, the fact that the point estimate of the relative effect is greater than 2.0, but the confidence interval is appreciably below 2.0, might make some hesitate in the decision to rate up for a large effect.

not follow the approach appropriately. First, ROBINS-I is currently the only available tool that explicitly includes a comparison against RCTs (and this situation is unlikely to change), and thus, this guidance only applies to the situations when ROBINS-I is used. Second, because of the advantage of assessing risk of bias on an absolute scale, use of ROBINS-I may facilitate combining results of RCTs and NRSs. However, under what conditions one should combine results from randomized and NRSs remain uncertain (this uncertainty also determines whether one uses ROBINS-I or other instruments for assessing risk of bias in NRS; Fig. 4).

Third, there is a possibility of misuse by those wanting to assign a higher certainty to a body of evidence from NRSs than is appropriate. Evaluators of evidence may rate the risk of bias from a group of NRSs as moderate risk of bias if they are not appropriately cautious about the impact of confounding and selection bias or if reporting is poor. Users of GRADE may then take the results of such studies and classify them as moderate certainty (if no problems in other GRADE domains exist) when, following current GRADE guidance, they should be classified as low certainty. This may result from a higher threshold and requirement for documenting upgrading rather than a potentially higher threshold for rating down NRS when using ROBINS-I. Fourth, furthermore, detailed guidance is required for appropriate application of ROBINS-I with more examples as concerns about the amount of time required and the lack of detailed reporting of risk of bias related items in current NRSs exist. Fifth, until now, there is no practical example to base a rating of initial high or even moderate certainty in the evidence that comes from a body of evidence from NRSs where no traditional GRADE-upgrading domain applies.

8. Unresolved issues

GRADE recognizes that there are a number of unresolved issues related to the arrival of ROBINS-I. The GRADE working group is aiming to address those in the near future. The unresolved issues are as follows:

1. If systematic review authors use ROBINS-I, should the results from NRSs and RCTs be considered together, including potentially in a meta-analysis (Fig. 4)? If RCTs and NRSs are indeed considered together, when should they be combined? Should NRSs be used to provide more precise estimates in summary effects when in fact NRS may dominate such estimates? Should they be used to alleviate concerns about indirectness because they often include broader populations and more practice oriented interventions? Should we continue to follow GRADE guidance to generally separate randomized and nonrandomized study results in GRADE summary tables such as evidence profiles and Summary of Findings (SoF) tables or should the guidance be modified? Until clear advice on when to combine data from randomized studies and NRS is available, we suggest following current GRADE guidance; if certainty of evidence differs in a body of randomized trials and a body of NRSs, one needs to only present SoF tables, the higher certainty evidence (almost invariable than that from RCTs). If certainty ratings are the same (typically low certainty), one presents results from the two bodies of evidence separately. If the results are consistent, then the overall certainty assessment is that of the two bodies of evidence (typically low certainty). If the results are inconsistent, and one believes both the bodies of evidence should be taken into consideration, then one will rate down further for this inconsistency, and the final rating will be one category lower (typically very low certainty).

2. How should we deal with publication bias in the context of including NRSs clearly posing more challenges than evaluating publication bias in RCTs (available evidence suggests that publication bias is a greater problem in NRSs than in RCTs)?

3. Under what circumstances should evidence syntheses broaden their scope of search and consider NRSs routinely?

4. GRADE needs to develop more detailed guidance than that currently exists regarding the presence of large effects and dose-effect relations. With regard to large effects, if a body of evidence from NRSs is indeed rated as high certainty in the evidence before the consideration of size of effect and very large effects exists, no further rating up is possible or required. This is also the situation regarding how GRADE currently deals with large effects observed in a body of evidence from RCTs. For instance, the large relative risk reduction observed with oral anticoagulation for the treatment of deep venous thrombosis for prevention of stroke in patients with atrial fibrillation does not lead to an ultimate rating of the certainty of the evidence beyond high.

5. Currently, GRADE has only three labels for risk of bias, such as not serious, serious, or very serious, on the risk of bias domain levels. For RCTs, this corresponds to rating of the body of evidence as high, moderate, or low certainty of evidence after considering risk of bias; for NRSs, this means that when GRADE currently uses serious risk of bias for NRSs, they are rated down from low initial certainty to very low. When raters use ROBINS-I with NRSs beginning at high certainty, three levels for rating down for risk of bias are required so that NRSs can arrive at a rating of very low certainty after considering risk of bias. GRADE is now exploring the best labeling options which include the use of not serious, serious,

---

Box 1 Clarification of terminology

GRADE uses the term "criteria" for all criteria in the evidence to decision frameworks of GRADE. Within these criteria, the "certainty in the evidence" (or quality or strength of evidence) is one criterion. Certainty of the evidence is assessed based on "certainty domains" with individual items within each domain. RoB is one domain; therefore, we will, in the context of GRADE, use the term RoB items to describe the seven areas of judgment that ROBINS-I calls domains.

---

very serious, and very, very serious leading to certainty ratings of high, moderate, low, and very low after risk of bias assessment.

9. Summary and next steps

Risk of bias can be best mitigated by a well-conducted RCT that balances known and unknown confounders, using the Cochrane RoB 2.0 tool or similar assessment tools for RCTs to assess risk of bias. For situations in which NRSs are used instead or in addition to RCTs, the arrival of ROBINS-I poses a number of opportunities and challenges to summarizing RoB in GRADE and raises a need for clarification about how ROBINS-I and GRADE are used together. Given the inherent limitations of studies that do not use randomization, a body of evidence from NRSs should generally not lead to moderate or high certainty in the evidence in relation to risk of bias. Raters using GRADE should always consider confounding and selection bias as reasons for rating down a body of evidence, and this is achieved in the current GRADE by assigning an initial rating of low certainty.

For studies of interventions that are assessed with ROBINS-I in the context of GRADE, we suggest that an initial rating of high is used, with appropriate consideration of the impact of lack of randomization leading to rating down for risk of bias according to the ROBINS-I tool. In practice, this will generally lead to rating down by at least two levels to low or very low certainty for NRSs. However, for results with large effects or dose response or results in which inference is strengthened by the plausible biases that exist, the extent of rating down may be lowered. We have not identified bodies of evidence in which a ROBINS-I assessment alone leads to no rating down or rating down by only one level. We therefore invite users of ROBINS-I and those who produce SoF tables or evidence profiles to submit to the GRADE working group any examples of when they believe that NRSs without reasons for rating up warrant moderate or high certainty evidence.

References

[1] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[2] Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knotterus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:380−2.

[3] Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014;349:g5630.

[4] Schunemann HJ, Best D, Vist G, Oxman AD, Group GW. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;169:677−80.

[5] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. ACP J Club 2008;149:2.

[6] Spencer FA, Iorio A, You J, Murad MH, Schunemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. BMJ 2012;345:e7401.

[7] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.

[8] Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013;66:151−7.

[9] Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: clinical practice guidelines. BMJ 2016;353:i2089.

[10] Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ 2016;353:i2016.

[11] Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol 2016;76:89−98.

[12] Burford BJ, Rehfuess E, Schunemann HJ, Akl EA, Waters E, Armstrong R, et al. Assessing evidence in public health: the added value of GRADE. J Public Health (Oxf) 2012;34:631−5.

[13] Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. Environ Int 2016;92-93:585−9.

[14] Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. PLoS Med 2007;4(5):e119.

[15] Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, 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.

[16] Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. J Clin Epidemiol 2017;87:14—22.

[17] Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, et al. GRADE: assessing the quality of evidence in environmental and occupational health. Environ Int 2016;92-93:611—6.

[18] Anderson STI, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36: 666—76.

[19] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.

[20] Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Community Health 2015;69: 189—95.

[21] Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

[22] Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. Annu Rev Public Health 2017;38: 39—56.

[23] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311—6.

[24] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017;87:4—13.

[25] Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schunemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 2002;166:1399—406.