Article published online: 2021-08-09

REVIEW ARTICLE

Are these results trustworthy? A guide for reading the medical literature

Fares Alahdab, Allison Morrow, Mouaz Alsawas, M. Hassan Murad

Evidence-based Practice Center, Mayo Clinic, Rochester, MN, USA

Access this article online

Website: www.avicennajmed.com

DOI: 10.4103/2231-0770.203611

Quick Response Code:

ABSTRACT

Physicians practicing evidence-based medicine need to be able to appraise a new study and determine whether the results warrant sufficient certainty to the level that they can be applied to patient care. Without such appraisal, misleading results can be incorporated into patient care, which can lead to inefficient, costly, and possibly harmful care. The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach offers a modern framework that can be applied to evaluate the trustworthiness of evidence. In this guide, we present a simplified approach based on GRADE; in which we call on readers of the medical literature to pay attention to six domains before making an overall judgment about the trustworthiness of results and before applying the evidence to patient care.

Key words: Assessment, critical appraisal, Development and Evaluation framework, evidence-based medicine, Grading of Recommendations, medical education

INTRODUCTION

Physicians practicing evidence-based medicine need to be able to appraise a new study and determine whether the results warrant sufficient certainty to the level that they can be applied to patient care. Without such appraisal, misleading results can be incorporated in patient care, which can lead to inefficient, costly, and possibly harmful care.

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach offers a modern framework that can be applied to evaluate the trustworthiness of evidence.[1] In this guide, we present a simplified approach based on GRADE; in which we call on readers of the medical literature to pay attention to six domains before making an overall judgment about the trustworthiness of results and before applying the evidence to patient care.

WHAT ARE WE RATING?

Trustworthiness of evidence is a complex construct that is intuitive (most people are able to understand) but yet is difficult to define. Various terms have been used to describe this construct (quality of evidence, strength of evidence, confidence in the estimates, and certainty in the evidence). In essence, we are trying to answer the question: Do I sufficiently trust this evidence to the extent that I can act on it and apply it to my patient?

A SINGLE STUDY OR A BODY OF EVIDENCE?

Ideally, patient care decisions should be based on a body of evidence.[2] This means a summary of all the relevant studies that have been systematically selected and appraised. The summary should provide estimates of effect for the outcomes of interest. Therefore, evaluating the extent of trustworthiness of evidence is optimally done when we have a systematic review. Other preappraised sources of evidence such as guidelines and synopses (e.g., ACP Journal Club)

Address for correspondence: Dr. Fares Alahdab, Mayo Clinic, Rochester, MN, USA.
E-mail: alahdab.fares@mayo.edu

Cite this article as: Alahdab F, Morrow A, Alsawas M, Murad MH. Are these results trustworthy? A guide for reading the medical literature. Avicenna J Med 2017;7:46-50.

are most helpful. Yet, clinicians encounter single studies published every day and they can apply the domains presented in this guide. Therefore, this guide should be applied optimally to a systematic review but can also be applied to individual studies.

In this guide, we present the six domains in the form of questions [Box 1]. We follow the description of each domain with examples and a tip (pearl) relevant to that domain.

WHAT IS THE EXTENT OF BIAS THAT MAY AFFECT EACH OUTCOME?

A basic question is how does bias influence the estimates of effect of each outcome? There are tools designed to evaluate the risk of bias based on each study design. For example, the Cochrane risk of bias for randomized trials,[3] the ROBINS tool for nonrandomized interventional studies[4] or the Newcastle–Ottawa tool for cohort and case–control studies,[5] and the QUADAS-2 for diagnostic studies.[6] For a clinician, applying these tools may be challenging as they require time and expertise. Therefore, clinicians may be better off focusing on the key 2–3 issues that would make them suspect bias. For example, was there a large loss to follow-up in a study? In a study of subjective outcomes (i.e., pain or quality of life), were the outcome assessors blinded? Did the two groups being compared have similar baseline characteristics?

Example

Two trials evaluated treatments for anxiety disorder in children. In one trial,[7] 90% of the 439 participants completed the final assessments, which implies a small loss to follow-up. In the second trial,[8] 16% of the participants dropped out before the final assessment, resulting in a large loss to follow-up. The first trial is at low risk of bias, whereas the second is at high risk.

Tip

When comparing study groups to ascertain baseline balance of prognostic factors (i.e., what is called), do not depend on the *P* value for the difference. This *P* value is meaningless because the study was never powered for that variable. A very large study can show statistically significant difference that is clinically trivial. It is more important to evaluate the absolute difference between the two groups in terms of that prognostic factor and make a judgment whether it is sufficient to bias the results.

Box 1: Questions that can help establish a level of trustworthiness in evidence*

What is the extent of bias that may affect each outcome?
Are the results precise?
Are the results consistent across studies?
Does the study directly answer my question?
Does it seem that reporting bias is a problem?
Are there any unusual factors making the observed association stronger (such as large effect, dose-response gradient, or confounders that strengthen the association)?

*Questions are best applied to a systematic review but can also be applied to a single study

ARE THE RESULTS PRECISE?

Precise results mean that our action would be the same if either boundary of the confidence interval were to represent the truth. If the confidence interval includes appreciable benefit and harm, the results are imprecise and we are uncertain about how to apply evidence to patient care. Furthermore, pay attention to the magnitude of the absolute difference or absolute change caused by the intervention. It is common to see a small or trivial effect that is statistically significant or has a large relative effect (i.e., relative risk [RR], hazard ratio or odds ratio [OR]), yet it is minor and not clinically important.

Example of adequate precision

Sersté *et al*.[9] reported an RR of all-cause mortality of 1.78 (1.36, 2.33). Both the lower and the higher limits of the 95% confidence interval show that there is higher mortality among patients with cirrhosis and ascites taking nonspecific beta-blockers (NSBB) compared to similar patients taking control medication.

Example of inadequate precision

In another study of similar patient sample, intervention, and outcome of Serste *et al*., Lo *et al*.[10] reported an RR of all-cause mortality of 0.98 (0.36, 2.65). This confidence interval is wider and it includes the RR value of 1, which means there is no difference between the two groups in terms of the outcome being evaluated. The lower limit of the 95% CI is on the side showing lower mortality among patients with cirrhosis and ascites taking NSBBs compared to controls. While the higher limit is on the side showing higher mortality among these patients. Therefore, there is high imprecision in this study for this particular outcome.

In a meta-analysis combining these two studies, as well as other studies that evaluated all-cause mortality among patients with cirrhosis and ascites taking NSBBs,[11] the pooled RR of 0.95 (0.67, 1.35) was also imprecise.

Tip

Do not pay too much attention to power and sample size calculation. These concepts are important when a study is being designed and are often manipulated by changing the delta (the difference in the outcome that is considered important and used for this calculation). Rather, when the results are available, focus on the confidence interval. Another tip, studies with more than 300 events usually lead to precise results.

ARE THE RESULTS CONSISTENT ACROSS STUDIES?

It is difficult to answer this question when reading a single study. The reader would need to have knowledge of the condition being studied and be aware of prior studies to determine if the new study provides consistent evidence. One can also evaluate if the effect is consistent across multiple subgroups in a single study. Inconsistency will lead to lower trust and lower certainty in the new study. When reading a meta-analysis, this question is much easier to answer because the various studies are summarized and plotted which allows visual and statistical evaluation of consistency.

Example of consistent results

Noori *et al.*[12] evaluated the development of retinopathy of prematurity among neonates who had lower versus higher O_2 saturation targets. They reported an RR of 0.31 (0.20, 0.48). This finding is in agreement with most of the studies on this topic, as evident by a meta-analysis[13] published recently.

Example of inconsistent results

In the previous example of the meta-analysis of NSBB and survival in patients with cirrhosis and ascites,[11] the studies on this topic were inconsistent with each other. We can see from the forest plot of all-cause mortality that some studies show benefit of NSBB in these patients (to the left of the line of no difference), some studies show harm (i.e., higher mortality) in patients taking NSBB (to the right of the line of no difference), and other studies show no difference at all (exactly on the line of no difference).

Tip

In a meta-analysis, consistent results are demonstrated visually in a forest plot by point estimates (results of individual studies) that are close to each other and with confidence intervals that overlap.[14] Statistically, consistent results will have high *P* value for heterogeneity test and low I-squared statistic.[14]

DOES THE STUDY DIRECTLY ANSWER MY QUESTION?

A patient may present asking for a diabetes medicine that lowers their risk of dying from a heart attack or of developing end-stage renal failure. If we read a study that only shows the effect of a diabetes drug on hemoglobin A1c (called a surrogate outcome), this study does not really answer our question about the patient-important outcomes. The study needs to have a similar patient, intervention and outcome, to the patient at hand. Otherwise, this evidence is less trustworthy because of "indirectness."

Example of indirect evidence

In Aversa *et al.*,[15] they evaluated the effect of testosterone on modifying cardiovascular risk factors and atherosclerosis progression in patients with metabolic syndrome and hypogonadism.

They evaluated surrogate markers of atherosclerosis such as high-sensitivity C-reactive protein and carotid intima-media thickness. These results are less trustworthy because patients are not interested in these laboratory and radiographic values, but rather care more about cardiovascular events.

Example of direct evidence

Araujo *et al.*[16] evaluated mortality outcomes and sex steroid levels. Vital status and mortality data were taken from the national death index. This is direct measurement of an outcome.

Tip

One other source of indirectness of evidence that may not be obvious relates to the study design. If we are looking for long-term outcomes and only found a short-term one; or if we are looking for a comparative effectiveness study (i.e., a study that compared two drugs head-to-head) and only found a study that compares drugs against placebo; the evidence is indirect.

DOES IT SEEM THAT REPORTING BIAS IS A PROBLEM?

Studies with positive results (i.e., statistically significant results) are more likely to be published (publication bias). Within a single study, outcomes that are statistically significant are more likely to be reported. If we suspect such reporting bias, we will clearly lose confidence in the results.

Example of publication bias

While conducting a systematic review evaluating the antidepressant reboxetine, it was found that data on 74% of patients enrolled in the trials were unpublished. Published data overestimated the benefit of the drug by 115% and underestimated harm.[17]

Tip

When reading a meta-analysis, you may encounter a funnel plot or statistical analysis to evaluate for publication bias. This analysis may only be reliable when you have more than twenty studies included in that particular analysis.[18]

ARE THERE ANY UNUSUAL FACTORS MAKING THE OBSERVED ASSOCIATION STRONGER?

There are three scenarios in which GRADE allows increasing certainty in evidence derived from observational studies.[19] These include a large effect size (i.e., a strong association, for example, RR >2 or <0.5); a dose–response gradient (i.e., the more of the intervention is given, the larger the effect is); and the presence of confounding that has an opposite direction of the traditional confounding (i.e., confounding that would strengthen the association). These scenarios are not common.

Example of a large effect

Evidence from observational studies[20] on infants sleeping position and sudden infant death syndrome (SIDS) found an OR of 4.1 (3.1, 5.5) of SIDS with front versus back sleeping positions. This led to the strong recommendations to put babies to sleep on their backs. Although this evidence came from observational studies, the large magnitude of effect warrants rating up the quality of evidence at least one level.

Example of dose–response gradient

The evidence from observational studies[21] on risk of bleeding in patients who are taking anticoagulation medications shows a dose-response relationship. The more the blood is thinned, the more the bleeding rate was. This leads us to increase our confidence in the results of these studies despite the fact that they are observational studies.

Example of confounding that strengthens the association

In a large systematic review of observational studies,[22] the findings showed higher death rates in private for-profit hospitals compared to private non-for-profit hospitals. The disease severity of patients admitted to the two hospitals is likely different, because sicker patients would tend to be admitted to the non-for-profit hospitals. This confounding effect would put the not-for-profit hospitals at a disadvantage, yet they are still showing lower mortality rates. This phenomenon would lead us to increase our confidence in this evidence.

Tip

It is important before we "raise our confidence" in a study that we make sure that the study has no important shortcomings. For example, if a study at high risk of bias shows a large effect, the bias may explain the large effect. Therefore, we should not raise our confidence in this situation.

SUMMARY

We presented six questions that are consistent with GRADE domains to help readers of medical literature come up with a global judgment about the trustworthiness of evidence. GRADE uses a semi-quantitative scale in which these domains can be added to an initial level of certainty (high, derived from randomized trials; or low, derived from nonrandomized studies) to reach a final judgment. In this guide, however, we hope that these six questions will help a reader without deep knowledge of methodology come up with an intuitive global judgment. This judgment will allow them to decide whether the study they are reading provides evidence that warrants sufficient certainty to be applied to their practice.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. 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.
2. Murad MH, Montori VM. Synthesizing evidence: Shifting the focus from individual studies to the body of evidence. JAMA 2013;309:2217-8.
3. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, *et al.* The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
4. Sterne JA, Hernán 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.
5. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, *et al.* The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analyses. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. [Last accessed on 2016 Nov 10].
6. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.* QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
7. Ginsburg GS, Kendall PC, Sakolsky D, Compton SN, Piacentini J, Albano AM, *et al.* Remission after acute treatment in children and adolescents with anxiety disorders: Findings from the CAMS. J Consult Clin Psychol 2011;79:806-13.
8. Melfsen S, Kühnemund M, Schwieger J, Warnke A, Stadler C, Poustka F, *et al.* Cognitive behavioral therapy of socially phobic children focusing on cognition: A randomised wait-list control study. Child Adolesc Psychiatry Ment Health 2011;5:5.
9. Sersté T, Melot C, Francoz C, Durand F, Rautou PE, Valla D, *et al.* Deleterious effects of beta-blockers on survival in patients with cirrhosis and refractory ascites. Hepatology 2010;52:1017-22.
10. Lo GH, Chen WC, Chen MH, Lin CP, Lo CC, Hsu PI, *et al.* Endoscopic ligation vs. nadolol in the prevention of first variceal bleeding in patients with cirrhosis. Gastrointest Endosc 2004;59:333-8.
11. Chirapongsathorn S, Valentin N, Alahdab F, Krittanawong C, Erwin PJ, Murad MH, *et al.* Nonselective ß-blockers and survival in patients

with cirrhosis and ascites: A systematic review and meta-analysis. Clin Gastroenterol Hepatol 2016;14:1096-104.e9.
12. Noori S, Patel D, Friedlich P, Siassi B, Seri I, Ramanathan R. Effects of low oxygen saturation limits on the ductus arteriosus in extremely low birth weight infants. J Perinatol 2009;29:553-7.
13. Fang JL, Sorita A, Carey WA, Colby CE, Murad MH, Alahdab F. Interventions to prevent retinopathy of prematurity: A meta-analysis. Pediatrics 2016;137. pii: E20153387.
14. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, *et al*. How to read a systematic review and meta-analysis and apply the results to patient care: Users' guides to the medical literature. JAMA 2014;312:171-9.
15. Aversa A, Bruzziches R, Francomano D, Rosano G, Isidori AM, Lenzi A, *et al*. Effects of testosterone undecanoate on cardiovascular risk factors and atherosclerosis in middle-aged men with late-onset hypogonadism and metabolic syndrome: Results from a 24-month, randomized, double-blind, placebo-controlled study. J Sex Med 2010;7:3495-503.
16. Araujo AB, Kupelian V, Page ST, Handelsman DJ, Bremner WJ, McKinlay JB. Sex steroids and all-cause and cause-specific mortality in men. Arch Intern Med 2007;167:1252-60.
17. Eyding D, Lelgemann M, Grouven U, Härter M, Kromp M, Kaiser T, *et al*. Reboxetine for acute treatment of major depression: Systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. BMJ 2010;341:c4737.
18. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006;333:597-600.
19. 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.
20. Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and the sudden infant death syndrome: Systematic review of observational studies and historical review of recommendations from 1940 to 2002. Int J Epidemiol 2005;34:874-87.
21. Schulman S, Beyth RJ, Kearon C, Levine MN; American College of Chest Physicians. Hemorrhagic complications of anticoagulant and thrombolytic treatment: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest 2008;133 6 Suppl: 257S-98S.
22. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann 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.