The American Journal of Surgery 220 (2020) 290–293

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

My Thoughts/My Surgical Practice

# The GRADE approach to appraising the evidence or how to increase the credibility of your research

One of the early meta-analyses on the timing of laparoscopic cholecystectomy after an episode of acute cholecystitis, published in 2006, included 4 studies with a cumulative number of 504 participants randomized to early or delayed laparoscopic cholecystectomy.[1] The authors found similar perioperative outcomes and a shorter hospital stay in patients undergoing early cholecystectomy. Since this publication, a number of cohort studies, randomized controlled trials (RCTs) and meta-analyses have found similar results and, additionally, have demonstrated a lower risk of wound infection following early cholecystectomy.[2–7] Interestingly, an analysis of the Hospital Episode Statistics in England found that, over the period from 1998 to 2012, only 16% of patients presenting with acute cholecystitis underwent early (within the same hospital admission) cholecystectomy, and no increase in the rate of early cholecystectomy was noted over time.[8] Anecdotal evidence suggests similar practice patterns in other European countries. Since high quality clinical research suggested that early cholecystectomy is superior to delayed surgery in at least some of the parameters of interest, how can the lack of transfer of research findings into clinical practice be explained?

This is a question the surgical community commonly faces. Journal editors are called to assess the credibility of clinical research and its potential to change clinical practice. Clinicians want to ascertain that research findings can be applied in their everyday practice and research evidence can translate into improved clinical outcomes and enhanced patient care. Guideline development bodies and policymakers face the challenge of appraising research evidence in the context of the so-called “real world”.

A significant scientific endeavor over the past decades has aimed at developing criteria to assess the certainty of clinical research evidence. The GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach is perhaps the most rigorously developed and most widely used methodology.[9,10] As such, it is endorsed and used by major organizations, including the World Health Organization, the Cochrane Collaboration and the National Institute for Health and Care Excellence.[11–13]

The GRADE working group has developed a framework of predefined parameters to appraise research evidence. It sets a starting point of high certainty of evidence (also called quality of evidence) for randomized trials and low certainty of evidence for observational studies. A number of parameters outlined hereafter downgrade or, occasionally, upgrade the certainty of evidence provided by a single study or a set of studies with similar PICO (Patient, Intervention, Comparator, Outcome) framework. As such, a low quality randomized clinical trial may be judged to provide very low certainty of evidence, whereas a high quality observational study can, under circumstances, provide high certainty of evidence. Following appraisal of the evidence, quality ratings for each outcome are specified as high, moderate, low or very low.

The following parameters are the pillars of GRADE, which define the quality of evidence and determine whether it should be upgraded or downgraded:

- **Study design:** Does the evidence come from randomized trials or observational studies?

This first criterion examines primarily the possibility of confounding bias affecting the comparative effect estimate. A randomized study design controls for both known and unknown factors that may be associated with both the exposure (intervention under investigation) and the outcome. In the aforementioned example, the randomized study design of the trials included in the meta-analysis precludes the possibility of patients with mild cholecystitis being selected for delayed cholecystectomy, thereby being at increased risk of developing perioperative complications due to fibrosis. Several other sources of bias are associated with observational evidence.[14]

- **Risk of bias:** What is the risk of bias of the study or studies under question?

Risk of bias assessment investigates the possibility of the conducted research having one or more systematic errors that might affect the occurrence or detection of outcomes of interest. For example, in the 2006 systematic review, none of the trials disclosed whether outcome assessors were blinded to the intervention. There are various instruments designed to assess the risk of bias. Typically, the Cochrane Tool or the Jadad score are used for randomized trials, whereas the ROBINS I tool or the Newcastle-Ottawa scale are used for observational studies.[15–18]

- **Inconsistency:** Are outcomes consistent across different studies?

When a number of studies are in agreement with each other with regard to the comparative effect of the interventions of interest, we can be more certain that this effect is true and not due to chance. Statistical inconsistency or heterogeneity is best assessed in the context of a meta-analysis (by calculating the tau-square and the I-square statistics). On the other hand, conceptual (clinical

https://doi.org/10.1016/j.amjsurg.2020.01.021

**Table 1**
GRADE parameters that define certainty of evidence with explanation. For further reading, see the GRADE Handbook.[19]

| | |
|---|---|
| Study design | Randomized versus observational study design. Evidence from randomized trials provide higher certainty in the evidence compared to observational studies. |
| **Risk of bias** | Risk of bias *across studies*, assessed with the Cochrane tool, ROBINS I or other instruments. High risk of bias undermines certainty of evidence. |
| **Inconsistency** | The magnitude of heterogeneity (variation in effect estimates among studies that cannot be explained by chance alone). This should be a summary assessment of statistical, clinical and methodological heterogeneity. This parameter cannot be assessed when a single study is appraised. Low inconsistency (lack of conflicting results) increases the certainty of evidence. |
| **Indirectness** | The extent to which findings across studies are applicable to our target population (or the average patient). Assessment of this parameter depends on whether the clinical setting, the target patient characteristics and the interventions are similar to those described in the studies. Significant indirectness will downgrade the certainty of evidence. |
| **Imprecision** | The degree of preciseness of the effect estimates, typically expressed by the 95% confidence interval. A narrow confidence interval suggests minimal imprecision and increases the certainty of evidence. |
| **Publication bias** | The potential effect of omitting non-published evidence, typically appraised visually with a funnel plot and statistically by using the Egger's or Begg's test. This parameter cannot be assessed when a less than 10 studies are appraised. High risk of publication bias will downgrade the certainty of evidence. |
| **Large effect** | If the magnitude of effect is high (an arbitrary cut-off is an upper or lower boundary of the risk ratio confidence interval >2 or <0.5, respectively) in the absence of evidence of confounding effects, then residual confounding is unlikely to change the direction of effects. A large effect will upgrade the certainty of evidence. Upgrading should rarely be used, only in the absence of known confounders. This parameter is only applicable to observational studies. |
| **Plausible confounding** | When residual confounding (confounding that has not been accounted for in adjusted analyses) would be expected to favor Intervention A over Intervention B, however effect estimates favor Intervention B or do not suggest difference in effects, the certainty of evidence may be upgraded. This parameter is only applicable to observational studies. |

and/or methodological) heterogeneity refers to whether study participant characteristics, interventions (or controls), and methods for outcome assessment are similar across studies. In our example, the proportion of patients with mild or moderate cholecystitis was not disclosed in any of the studies, whereas baseline patient demographics were not reported in 3 out of the 4 RCTs, thereby not allowing a proper assessment of clinical heterogeneity. Inspection of the forest plot and an $I^2$ value of 40% in the example meta-analysis suggest moderate heterogeneity, which could be attributed to clinical heterogeneity.

- **Indirectness:** Does the study framework apply to my practice/to the average practice?

When appraising research evidence, we need to consider the applicability of research findings in the context of interest – the concept of "directness". For instance, patients recruited in the 4 RCTs of the example meta-analysis were admitted in the hospital with documented gallbladder inflammation, but it is unclear how many patients suffered from mild, moderate or severe cholecystitis. The putative similar effect of early and delayed cholecystectomy on postoperative outcomes may not be applicable to moderate

**Table 2**
GRADE evidence profile: Early versus delayed cholecystectomy for acute cholecystitis. Based on: *Lau* et al. *Surg Endosc 2006;20:82-87.* Created using GRADEpro GDT: GRADEpro Guideline Development Tool [Software]. McMaster University, 2015 (developed by Evidence Prime, Inc.). Available from gradepro.org.

| Certainty assessment | | | | | | | № of patients | | Effect | | Certainty | Importance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| № of studies | Study design | Risk of bias | Inconsistency | Indirectness | Imprecision | Other considerations | early | delayed cholecystectomy | Relative (95% CI) | Absolute (95% CI) | | |
| Complications | | | | | | | | | | | | |
| 4 | randomised trials | serious[a] | serious[b] | serious[c] | not serious | none | 36/254 (14.2%) | 35/237 (14.8%) | **OR 0.97** (0.59 −1.61) | **4 fewer per 1.000** (from 55 fewer to 70 more) | ⊕◯◯◯ VERY LOW | IMPORTANT |
| **Intra-abdominal collection** | | | | | | | | | | | | |
| 4 | randomised trials | serious[a] | not serious | serious[c,d] | very serious[e] | none | 11/254 (4.3%) | 8/237 (3.4%) | **OR 1.28** (0.51 −3.25) | **9 more per 1.000** (from 16 fewer to 68 more) | ⊕◯◯◯ VERY LOW | IMPORTANT |
| **Bile leakage** | | | | | | | | | | | | |
| 4 | randomised trials | serious[a] | serious[f] | serious[c,d] | very serious[e] | none | 7/254 (2.8%) | 2/237 (0.8%) | **OR 2.22** (0.64 −7.72) | **10 more per 1.000** (from 3 fewer to 53 more) | ⊕◯◯◯ VERY LOW | IMPORTANT |
| **Length of hospital stay** | | | | | | | | | | | | |
| 3 | randomised trials | serious[a] | serious[g] | serious[c] | not serious | none | 180 | 179 | – | MD **1.14 days fewer** (1.58 fewer to 0.7 fewer) | ⊕◯◯◯ VERY LOW | NOT IMPORTANT |

**CI:** Confidence interval; **OR:** Odds ratio; **MD:** Mean difference.
Explanations.
[a] Due to unknown risk of outcome assessment bias.
[b] Poor reporting of baseline parameters across RCTs to allow assessment of clinical heterogeneity. Considerable statistical heterogeneity ($I^2 = 40\%$).
[c] Unclear proportion of patients with mild versus moderate disease. Poor reporting of baseline patient characteristics.
[d] No reporting of method for outcome assessment.
[e] Very wide interval estimate.
[f] Poor reporting of baseline parameters across RCTs to allow assessment of clinical heterogeneity. Considerable statistical heterogeneity ($I^2 = 52\%$).
[g] No information to calculate $I^2$ was provided by the review authors.

cholecystitis, if the majority of patients enrolled in those trials had mild disease. In addition, differences between the studies and "real world" practice in patient characteristics or perioperative management might also influence our judgement on indirectness. For instance, if the authors of the sample RCTs had considered only young and fit patients, the outcomes would probably not be applicable to the average patient admitted with acute cholecystitis in a given setting. External validity also refers to potential variability in infrastructure, logistics, resources, and surgical team expertise in different clinical settings within or across different countries.

- **Imprecision:** How certain are we about the relative effect of an intervention?

The lack of statistical significance does not necessarily mean that there is lack of effect. The certainty or uncertainty about the comparative effect is reflected on the width of the confidence interval of the effect estimate. In our example meta-analysis, the confidence interval of the odds ratio for the outcome "complications" is 0.59–1.61, suggesting that the true point estimate of effect probably lies within a relatively narrow range. We can therefore be quite confident about the lack of difference in the summary effects across the RCTs considered in the meta-analysis.

However, the finding of statistical significance does not necessarily imply clinical significance. For example, the confidence interval for the outcome hospital stay was −1.58 to −0.70 days, which means that early cholecystectomy was associated with a decreased duration of hospital stay by approximately 0.5–1.5 days. Although this difference is statistically significant, its clinical significance may be considered marginal.

- **Other parameters:** Publication bias, confounding effect, large effect

The GRADE approach considers a number of other factors that may influence the certainty of evidence. The presence of publication bias suggests that failure to include certain publications (usually small studies) might have affected effect estimates. Large magnitude of effect (as determined by a relative risk of >2, in the absence of known confounders in observational studies) and a dose-response gradient (in pharmacological studies) would also increase our certainty in the evidence.

It is obvious that risk of bias (the methodological quality of a study or a set of studies) and the study design (randomized or observational) are not the only determinants of the certainty in evidence.

The above considerations might seem somewhat complex: how do these parameters determine the certainty of evidence? An outline of the methodology is presented in Table 1, whereas a detailed guideline on how to apply the GRADE approach and when to upgrade or downgrade the certainty of evidence is provided in a series of articles published by the GRADE working group and summarized in the GRADE Handbook.[19] Fortunately, advanced web-based platforms simplify this task, facilitating application of the GRADE methodology and automatically calculating the certainty of evidence. MAGICapp and GRADEpro are free online platforms that are increasingly used for the assessment of the certainty of evidence by authors of systematic reviews and guideline developers.[20,21]

It is noteworthy that the GRADE methodology can be applied in a set of studies (usually in the form of a meta-analysis) or in individual studies. It does not apply to basic or translational research. If one would apply the GRADE methodology to assess the certainty of evidence provided by our example meta-analysis, they would find that the quality of evidence across outcomes is *very low* (Table 2). This can partly explain why the findings of numerous cohort studies, RCTs and meta-analyses have not been transferred into clinical practice.

Taking it one step further, the GRADE approach suggests considering several parameters that might affect the decision to recommend for or against an intervention, which constitute the evidence-to-decision framework. Most importantly, the balance between desirable and undesirable effects, the certainty of the evidence, the resources required to implement the intervention, its feasibility and its acceptability to stakeholders may affect the decision to recommend an intervention. In the case of early cholecystectomy, even in the light of recent meta-analyses, plausible explanations for failure to transfer research evidence into clinical practice are the uncertainty of evidence, lack of clinically significant effects and poor clinical applicability due to limited human resources and healthcare infrastructures. The evidence-to-decision framework is formed by guideline developers to inform practice recommendations.

Authors of systematic reviews submitting their work to *The American Journal of Surgery* are strongly encouraged to apply the GRADE methodology, a powerful tool allowing evidence-based assessment of the certainty of evidence and its applicability to clinical practice. Such practice has the benefit of informing decision making with the ultimate goal of improved clinical outcomes and enhanced patient experience and care.

## References


1. Lau H, Lo CY, Patil NG, Yuen WK. Early versus delayed-interval laparoscopic cholecystectomy for acute cholecystitis: a metaanalysis. *Surg Endosc Other Interv Tech*. 2006. https://doi.org/10.1007/s00464-005-0100-2.
2. Casillas RA, Yegiyants S, Collins JC. Early laparoscopic cholecystectomy is the preferred management of acute cholecystitis. *Arch Surg*. 2008. https://doi.org/10.1001/archsurg.143.6.533.
3. Gutt CN, Encke J, Köninger J, et al. Acute cholecystitis: early versus delayed cholecystectomy, a multicenter randomized trial (ACDC Study, NCT00447304). *In: Ann Surg*. 2013. https://doi.org/10.1097/SLA.0b013e3182a1599b.
4. Özkardeş AB, Tokaç M, Dumlu EG, et al. Early versus delayed laparoscopic cholecystectomy for acute cholecystitis: a prospective, randomized study. *Int Surg*. 2014. https://doi.org/10.9738/INTSURG-D-13-00068.1.
5. Roulin D, Saadi A, Di Mare L, Demartines N, Halkic N. Early versus delayed cholecystectomy for acute cholecystitis, are the 72 hours still the rule? A randomized trial. *Ann Surg*. 2016. https://doi.org/10.1097/SLA.0000000000001886.
6. Wu XD, Tian X, Liu MM, Wu L, Zhao S, Zhao L. Meta-analysis comparing early versus delayed laparoscopic cholecystectomy for acute cholecystitis. *Br J Surg*. 2015. https://doi.org/10.1002/bjs.9886.
7. Lyu Y, Cheng Y, Wang B, Zhao S, Chen L. Early versus delayed laparoscopic cholecystectomy for acute cholecystitis: an up-to-date meta-analysis of randomized controlled trials. *Surg Endosc*. 2018. https://doi.org/10.1007/s00464-018-6400-0.
8. Murray AC, Markar S, Mackenzie H, et al. An observational study of the timing of surgery, use of laparoscopy and outcomes for acute cholecystitis in the USA and UK. *Surg Endosc*. 2018;32(7):3055–3063. https://doi.org/10.1007/s00464-017-6016-9.
9. Guyatt GH. *GRADE : An Emerging Consensus on Rating Quality of Evidence and Strength of Recommendations*. vol 336. 2008 (April).
10. Alhazzani W, Guyatt G. An overview of the GRADE approach and a peek at the future. *Med J Aust*. 2018;209(7):291–292. https://doi.org/10.5694/mja18.00012.
11. World Health Organization. *WHO Handbook for Guideline Development*; 2012. Published http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf. Accessed October 2, 2017.
12. Schünemann HJ, Oxman AD, Vist GE, et al. The GRADE approach. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*; 2011. Version 5. https://handbook-5-1.cochrane.org/chapter_12/12_2_1_the_grade_approach.htm
13. The guidelines manual | guidance and guidelines | NICE. https://www.nice.org.uk/process/pmg6/chapter/reviewing-the-evidence. Accessed October 2, 2017.
14. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002. https://doi.org/10.1016/S0140-6736(02)07451-2.
15. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011. https://doi.org/10.1136/bmj.d5928.
16. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Contr Clin Trials*. 1996. https://doi.org/10.1016/0197-2456(95)00134-4.

17. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016. https://doi.org/10.1136/bmj.i4919.
18. Wells G, Shea B, O'Connell D, Peterson J. The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analyses. Ottawa, ON: Ottawa Hospital Research Institute.
19. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group.
20. GRADE Working Group. *GRADEpro Guideline Development Tool [software]*. McMaster Univ; 2015.
21. MAGICapp. Available in: https://app.magicapp.org. Accessed October 31, 2019.

Stavros A. Antoniou*
*Medical School, European University Cyprus, Nicosia, Cyprus*

*Department of Surgery, Mediterranean Hospital of Cyprus, Limassol, Cyprus*

George A. Antoniou
*Department of Vascular & Endovascular Surgery, The Royal Oldham Hospital, Pennine Acute Hospitals NHS Trust, Manchester, United Kingdom*

*Division of Cardiovascular Sciences, School of Medical Sciences, University of Manchester, Manchester, United Kingdom*

* Corresponding author. Athinon-Souniou 11, 19001, Keratea Athens, Greece.
*E-mail address:* stavros.antoniou@hotmail.com (S.A. Antoniou).

12 January 2020