Techniques in Coloproctology (2019) 23:919–923
https://doi.org/10.1007/s10151-019-02069-9

How to find a meta-analysis you can trust

R. L. Nelson[1]

Received: 31 July 2019 / Accepted: 14 August 2019 / Published online: 28 August 2019

History

I first read of meta-analysis in 1981, in a paper by Baum et al. in the New England Journal [1]. It described a meta-analysis of 17 randomized trials assessing the value of prophylactic antibiotics in colorectal surgery—any antibiotics given by any route. They found a 75% reduction in surgical wound infection and pulled no punches in their conclusions. They wrote that it was now unethical to do colorectal surgery without prophylactic antibiotic cover. I found that very interesting, but, having just begun as an attending surgeon, I went on to other things and forgot it. Its senior author was one of the true fathers of this science [2].

Six years later, Richard Peto wrote why systematic review and meta-analysis were so important in modern clinical medicine [3]. I read that paper but again found it rather too easy to forget, until 11 years later when I was asked by the American Society of Colon and Rectal Surgeons (ASCRS) to join the Cochrane Collaboration in their behalf. The Cochrane Collaboration was formed in 1993 and the first edition of the Cochrane Library published in 1995. Putting this Library in its place, there are many astounding advances in medicine in the twentieth century, but this could be the greatest. It gave a tool to practicing physicians, insurers, governments that enabled them to use the very best evidence available to optimally treat a very large number of diseases. Meta-analysis passed from the epidemiological journals to practitioner's desk.

There were detractors, but no one could come up with a better alternative. Still problems with meta-analysis were discussed, most thoroughly within the methods' groups of the Cochrane Collaboration. As a result, the process has grown in complexity since the late twentieth century editions of the Cochrane library in an effort to shore up the leaks, those leaks being bias. Publication of systematic reviews with meta-analyses has also grown hugely in volume outside of the Cochrane library. As opposed to a randomized clinical trial, all that is required to do it is access to a medical library. Meta-analysis software, most of it being free, including Cochrane's, is copiously available. In many cases, such as antibiotic treatment of appendicitis, there are many more published systematic reviews and meta-analyses than there are eligible randomized trials. In 1981, three meta-analyses popped up in PubMed, and in 2017 there were 19,500. For any single medical topic meta-analyses differ greatly in the literature search, methodology and of course quality.

So is there any way to tell the wheat from the chaff? The authors discuss several things to look for: heterogeneity, multiplicity, risk of bias, publication bias, random error and sample size [4]. Ultimately, they hope that journals will become more discerning. That is easier said than done. Are there journals that reliably publish high quality reviews? The Cochrane library has been the gold standard for 24 years. However, even they have hit some bumps in the road in recent years [5]. In an effort to be thorough in avoiding bias and misinterpretation their reviews have become increasingly ponderous, long, repetitive, jargon laden and have been directed away from practicing physicians and toward funders and guideline organizers. Some have worried about the relevance of their reviews [6].

The solution

In any case, all the topics above have had thousands of pages written about them, and I have not even brought up "*p*" values [7]. But the best tool to evaluate a systematic review is not the forest plot, the odds ratio, the confidence intervals or the heterogeneity, but the actual overall quality of the evidence:

GRADE: The Grades of Recommendation, Assessment, Development and Evaluation (GRADE) approach can be used to classify the quality of evidence [8].

GRADE has five domains:

✉ R. L. Nelson
altohorn@uic.edu

[1] Epidemiology/Biometry, University of Illinois School of Public Health, 1603 West Taylor, Chicago, IL 60612, USA

1. Risk of bias.
2. Inconsistency (i.e., heterogeneity).
3. Indirectness (for instance, there may be various measures of the primary outcome, so perhaps fever for wound infection or adenoma recurrence for colon cancer risk).
4. Imprecision (wide confidence intervals or too small studies, or too few studies or too few outcome events. Random error dominates in this instance.)
5. Publication bias (or the small studies effect, i.e., small studies with insignificant results tend not to get published. The positive small studies which are published then bias the meta-analysis in favor of the intervention) [9].

In doing a GRADE assessment, for reviews of randomized trials, the quality at the beginning starts with a rating of "High quality of evidence". When serious (1 step down) or very serious (2 steps down) concerns arise for any one or more of the domains, GRADE moves down one, two or three places to Moderate, Poor, and Very Poor. If the quality of the evidence is Very Poor, it means that no matter how statistically significant the evidence in favor of an intervention, there is not enough evidence to recommend its usage.

There is software called GRADEPRO that was downloadable but now only available on line. However, the secret is that you do not need the software to do GRADE. It is just as accurate and much easier not to use the software.

Several things need to be mentioned about each of the domains.

Risk of bias: (RoB) also has five or so domains (Fig. 1) and others can and should be added that would be unique to the review being performed: e.g., duration of follow-up after treatment of anal fissure or whether a validated instrument was used for the assessment of anal incontinence in continence reviews.

Choice of allocation sequence:There are good ones like computer-generated random numbers, and horrible ones that are very prone to bias such as medical record numbers or days of the week.

Blinding: I have no idea what double blind means. Many individuals need to be blinded in a randomized trial from those dealing with the allocation concealment at the time of randomization to carers, to patients, and to outcome assessors.

Sample size: should be a calculation before data. Both GRADE and trials sequential analyses do post hoc assessments which is frowned upon [10].

Attrition: Everybody screws it up. Randomization is not baptism. It is very easy to break it apart. Once somebody is randomized, they are in their allocation group for life. The authors must describe in detail how they handled dropouts and an analysis ignoring the drop outs is deeply flawed. Some randomized studies report per protocol results. That is not a randomized study, but a case–control study, which

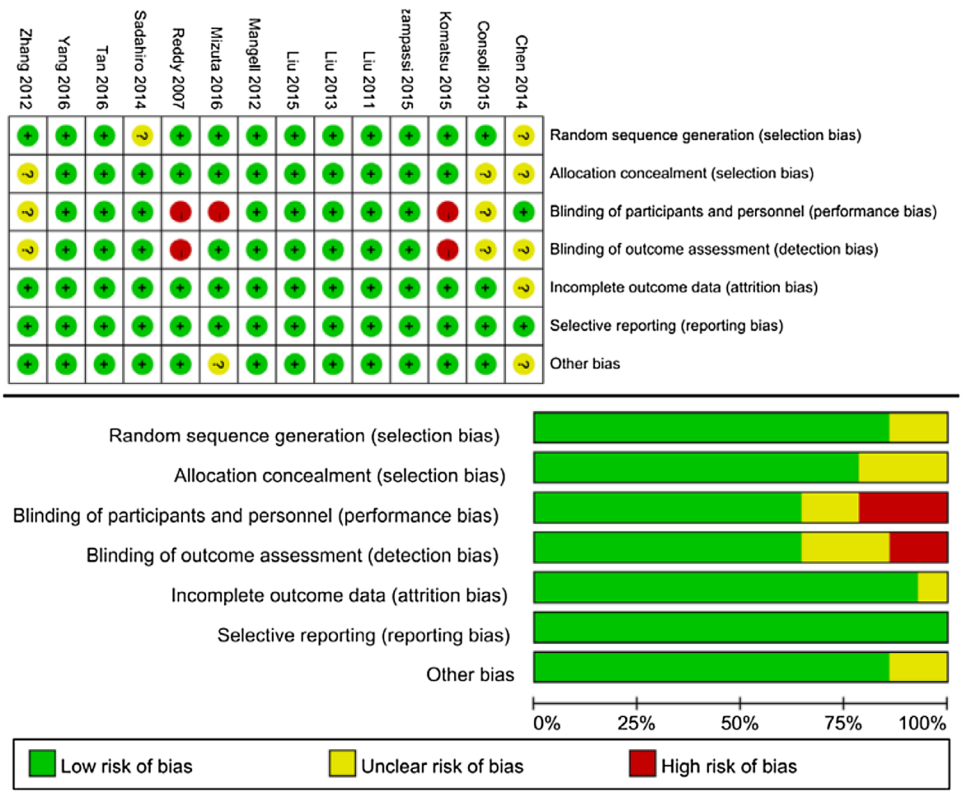

Fig. 1 Probiotics to prevent surgical wound infection in colorectal surgery. Risk of bias (Wu [11].)

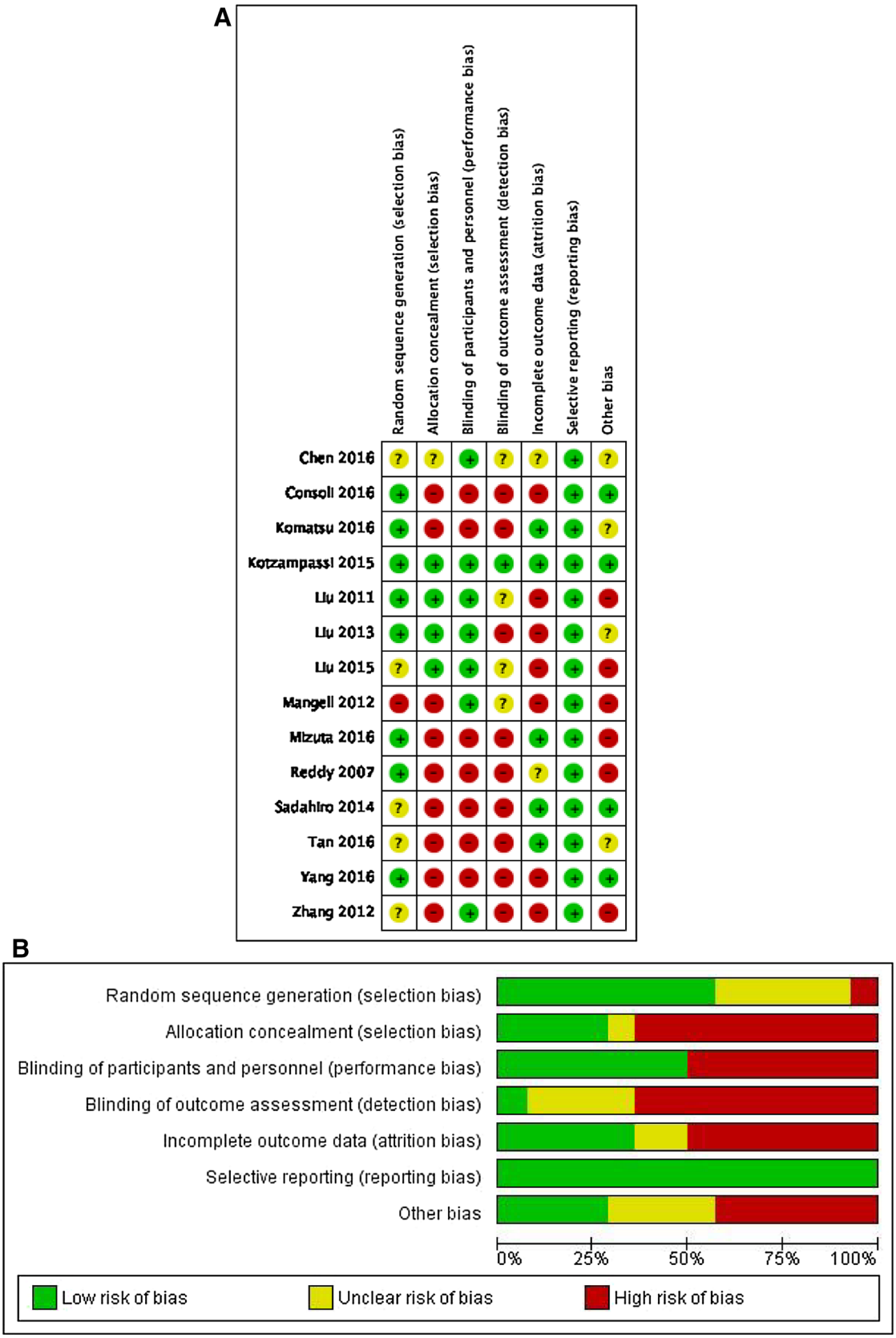

Fig. 2 Reassessment of risk of bias in each study included in Wu [1]

is very poor evidence. Or a modified intention to treat analysis may be reported, which again breaks the randomization by exclusion of randomized individuals. Exclusions for any reason are never made at random.

And so on. Study authors are often given the benefit of the doubt by reviewers in their presentations in published meta-analyses for no good reason. The authors already have their publication. Nobody can take that away from them. As a reviewer, the only way to read a paper is to assume that if the study authors did not write it, they did not do it; period. And so a red dot should be given in the RoB Figure if the mechanism of randomization is not specified (or at best, yellow), or if who exactly was blinded is not specified, etc. A very green RoB figure in a published meta-analysis, if the Cochrane software figure is used (Fig. 1), is a real danger sign. This is demonstrated in the comparison of Figs. 1 and 2. Figure 1 shows the risk of bias summary from an otherwise excellent paper studying the use of probiotics with antibiotics to prevent surgical wound infection in colorectal surgery. [11]. It is too green in the graphic, meaning that the RoBs for each risk in each study were judged to be low. Figure 2 shows this author's review of the same included studies, adhering to the principal of "If they did not write it, they did not do it". That would get a very serious mark for RoB in GRADE, down two points to "Poor" even before you get to the other domains. The less the RoB specificity by an author, with or without colored figures, the less reliable is the meta-analysis.

Lastly, there is one Cochrane editor who is convinced that many randomized trials never happened, but were totally dry labbed [12]. I know of no reliable way to detect these, though trial registries, like ClinicalTrials.gov should help prevent this.

Multiplicity: If a study author chooses a significance level of $p < 0.05$, and 20 outcomes are measured in the randomized trial, the odds of having a significant outcome for one among the 20 is about 100%. Not 0.05. Of course that one finding is what is trumpeted in the publication The more outcomes that are measured, the more the p value of significance must be lowered [13] by study authors to adjust for multiplicity. I have never seen that done in print. There is one new medication for C. difficile that is an excellent example of this bias [14].

Heterogeneity: Not exactly a bad thing, but it must be explored. What caused it? How successfully it is explored will determine how much GRADE is lowered for inconsistency. If a cause is not found and no adjustment made, GRADE must be lowered.

Indirectness:Surrogate outcomes make a lot of studies feasible (like diet and colon cancer risk using adenoma recurrence in place of cancer occurrence), but there are no good surrogates.

Imprecision: Trial sequential analysis and GRADE both deal with this [10]. There are way too many tiny randomized trials. They just muddy the waters. Traditionally Cochrane preached the inclusion of all studies. RCTs with 20 or 40 or 60 patients are pilot studies to support the doing of bigger studies and probably should be excluded if larger ones exist. If those larger ones do not exist, GRADE must be lowered.

Publication bias: Same thing. Insignificant small studies do not get published [9], so to balance this error, significant small studies should be excluded. I find funnel plots too difficult to interpret. I have never seen a GRADE marked down because of a funnel plot.

So that is how to find a "true" systematic review. Once you get used to seeing the red flags, it is not so difficult. As tools for medical practice, systematic review and meta-analysis are indispensable [15], providing the second deep look at medical innovations [16].

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Ethical approval This study used no patient information or protected health information and was exempt from Institutional Review Board approval.

Informed consent All survey participants were anonymous, no identifying information was collected, and all consented to participation in the project.

References

1. Baum ML, Anish DS, Chalmers TC, Sacks HS, Smith H Jr, Fagerstrom RM (1981) A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. N Engl J Med 305:795–799
2. Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R (1987) Meta-analysis of clinical trials as a scientific discipline. II: replicate variability and comparison of studies that agree and disagree. Stat Med 6(7):733–744
3. Peto R (1987) Why do we need systematic overviews of randomized trials? Stat Med 6(3):233–244
4. Doleman B, Williams JP, Lund J (2019) Why most published meta-analysis findings are false. Tech Coloproctol. https://doi.org/10.1007/s10151-019-02020-y
5. Newman M (2019) Clarification of news feature "Has Cochrane lost its way?". BMJ 364:l670
6. Ioannidis JPA (2019) Cochrane Crisis: secrecy, intolerance and evidence-based values. Eur J Clin Invest 49(3):e13058
7. Sterne JA, Smith GD (2001) Sifting the evidence-what's wrong with significance tests? BMJ 322(7280):226–231
8. Schünemann H, Brożek J, Guyatt G, Oxman A (eds) (2013) GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group

9. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr (1987) Publication bias and clinical trials. Control Clin Trials 8(4):343–353
10. Castellini G, Bruschettini M, Gianola S, Gluud C, Moja L (2018) Assessing imprecision in Cochrane systematic reviews: a comparison of GRADE and trial sequential analysis. Syst Rev 7(1):110
11. Wu XD, Xu W, Liu MM, Hu KJ, Sun YY, Yang XF, Zhu GQ, Wang ZW, Huang W (2018) Efficacy of prophylactic probiotics in combination with antibiotics versus antibiotics alone for colorectal surgery: a meta-analysis of randomized controlled trials. J Surg Oncol 117(7):1394–1404
12. Roberts I, Smith R, Evans S (2007) Doubts over head injury studies. BMJ 334(7590):392–394
13. Dmitrienko A, D'Agostino RB Sr (2018) Multiplicity considerations in clinical trials. N Engl J Med 378(22):2115–2122
14. Nelson RL, Suda KJ, Evans CT (2017) Antibiotic treatment for Clostridium difficile-associated diarrhoea in adults. Cochrane Database Syst Rev 3:CD004610. https://doi.org/10.1002/14651858.CD004610.pub5
15. Ioannidis JPA (2017) Meta-analyses can be credible and useful. JAMA Psychiatry 74(4):311–312
16. Ioannidis JPA (2018) Why replication has more scientific value than original discovery. Behav Brain Sci 41:e137