Citation: Clinical and Translational Gastroenterology (2017) 8, e93; doi:10.1038/ctg.2017.20
Official journal of the American College of Gastroenterology

www.nature.com/ctg

PRIMERS IN CLINICAL AND TRANSLATIONAL RESEARCH

How to Conduct and Interpret Systematic Reviews and Meta-Analyses

Siddharth Singh, MD, MS[1,2]

Systematic reviews with or without meta-analyses serve a key purpose in critically and objectively synthesizing all available evidence regarding a focused clinical question and can inform clinical practice and clinical guidelines. Performing a rigorous systematic review is multi-step process, which includes (a) identifying a well-defined focused clinically relevant question, (b) developing a detailed review protocol with strict inclusion and exclusion criteria, (c) systematic literature search of multiple databases and unpublished data, in consultation with a medical librarian, (d) meticulous study identification and (e) systematic data abstraction, by at least two sets of investigators independently, (f) risk of bias assessment, and (g) thoughtful quantitative synthesis through meta-analysis where relevant. Besides informing guidelines, credible systematic reviews and quality of evidence assessment can help identify key knowledge gaps for future studies.

Clinical and Translational Gastroenterology (2017) **8,** e93; doi:10.1038/ctg.2017.20; published online 18 May 2017

Well-designed, single studies are often proposed to be the ultimate answer to a clinical question, and positive findings, particularly from interventional studies, are viewed very favorably by the research enterprise, including funders, fellow researchers, journals, and the lay media. However, it is important to recognize that early (and often the most highly cited) studies in a field tend to overestimate or inflate magnitude of benefits, due to study design features (for example, inclusion of very-high-risk patients, use of composite end points, or surrogate end points) or overoptimistic sample size calculations (resulting in large, but imprecise, estimates).[1–3] Subsequent, independent, similarly designed studies often fail to show the large effect estimates seen with the first study, and with time, the “truth wears off”.[4,5] Hence, by examining the totality of evidence, rather than relying on individuals studies, we can improve the usefulness of results to clinical decision-makers, and can: (a) calibrate confidence in the estimates based on consistency with other studies, (b) improve precision of findings, (c) avoid premature closure about the magnitude of effect before the estimate has moderated through repeated and independent evaluation, and (d) prevent premature closure of a potentially effective intervention owing to concerns for non-significant results.[6]

One approach to understanding the body of evidence is through well-conducted systematic reviews with or without meta-analyses. In contrast to traditional, unstructured narrative reviews, which provide a broad overview of clinical and scientific developments in a particular field, systematic reviews address a focused clinical question in a structured and reproducible manner. It is often, but not always, accompanied with a meta-analysis, which is a statistical pooling of results from different studies to derive a single summary effect estimate, with more precision.[7] In this primer, we will discuss the steps involved in conducting and reporting a systematic review and meta-analysis, commonly observed mistakes and misunderstandings in their conduct and interpretation, and newer concepts, such as network meta-analysis and assessing quality of the entire body of evidence.

HOW TO CONDUCT A SYSTEMATIC REVIEW AND META-ANALYSIS AND COMMON PITFALLS TO AVOID

The key steps in designing and conducting a systematic review, as well as common pitfalls, are summarized in Table 1 and detailed below. Minimum evidence-based reporting items when conducting and reporting systematic reviews and meta-analysis, such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement may be used by investigators and journal reviewers, to critically appraise reporting of systematic reviews.[8]

Formulating a focused clinical question. The first (and in my experience, a very crucial step in determining whether or not a review would be completed to publication) is formulating a focused clinical question. This question often originates in one’s clinical practice, where there is equipoise in medical literature in choosing one intervention or diagnostic strategy over another, or identifying rates, risk factors, or prognostic factors in defined group of patients. This question should not be too broad (infeasible, and often at-risk for conceptual heterogeneity when considering meta-analysis; for example, what are the treatment options for Crohn’s disease) or too narrow (may not be too clinically relevant and generalizable,

[1]Division of Gastroenterology, University of California San Diego, La Jolla, California, USA and [2]Division of Biomedical Informatics, University of California San Diego, La Jolla, California, USA
Correspondence: S Singh, MD, MS, Division of Gastroenterology, University of California San Diego, 9452 Medical Center Drive, 1W 501, La Jolla, California 92093, USA. E-mail: sis040@ucsd.edu
Received 15 March 2017; accepted 5 April 2017

Table 1 Steps in designing and conducting systematic reviews and meta-analysis, and common pitfalls in conducting and interpreting meta-analysis

	Steps	Important elements	Common pitfalls
1	Develop a focused clinical question	Clinically relevant question, often derived from equipoise observed in literature and clinical practice	•Too broad or too narrow a question, which limit feasibility and relevance of a systematic review •Performing quantitative synthesis, without clear clinical basis (in absence of equipoise about direction or magnitude of benefit)
2	Developed systematic review protocol	Use PICO format and develop explicit inclusion and exclusion criteria and identify *a priori* hypothesis to explain anticipated heterogeneity	•Lack of clarity in scope and purpose, due to absence of *a priori*-defined patient population, outcomes of interest, and interventions •Non-specific subgroup analyses without meaningful hypotheses for why results would be inconsistent across studies
3	Systematic literature review	Engage medical librarian to conduct a sensitive search pertinent to PICO question, of multiple databases, including conference proceedings, clinical trial registries, and gray literature, as well as recursive search of systematic reviews	•Clinician-designed search, of single database, which are generally too specific, rather than being sensitive, which may miss several potential studies •Failure to search conference proceedings, trial registries, which may exacerbate the file-drawer problem •Search restricted to English language
4	Study identification	Screening titles and abstracts and full texts based on inclusion/exclusion criteria by two investigators independently	•Single investigator identifying studies, resulting in potentially missed studies or inappropriate inclusion/exclusion of studies
5	Data abstraction	Abstract all relevant study data, using piloted data abstraction form, by two investigators independently	•Single investigator abstracting data, without confirmation by another investigator, which may introduce bias or errors •Failure to anticipate and abstract data, which may be inconsistently reported across studies (for example, differences in definition of outcome, drug doses/schedules, co-interventions, etc.)
6	Risk of bias assessment	Critical assessment of study quality, through a combination of standardized tools and investigator-identified factors that may bias results, in duplicate, independently	•Failure to systematically and critically appraise quality of included studies •Strict adherence to elements reported only in risk of bias tools, without adaptation of focused clinical question, frequently resulting in failure to identify potential sources of bias •Use of quantitative scoring to stratify studies as “high” or “low” quality, failing to recognize that different elements are not weighted equally across clinical questions
7	Quantitative synthesis or meta-analysis	If appropriate (studies conceptually similar), perform meta-analysis, generally using random-effects model, estimating effect estimate and confidence intervals, statistical and conceptual assessment of heterogeneity, subgroup and sensitivity analyses, and small study effects	•Performing meta-analysis, even if studies are conceptually heterogeneous, findings from which are not applicable to clinical practice and may misinform lay audience •Using statistical measures of heterogeneity to determine whether fixed- or random-effects model should be adopted •Overinterpretation and inappropriate interpretation of subgroup analyses and failure to critically analyze and acknowledge causes for heterogeneity •Overinterpreting findings from small studies due to failure to recognize small study effects

PICO, Patients, Interventions, Comparators and Outcomes.

with only 1–2 studies to warrant a systematic review or meta-analysis; for example, what is the benefit of certolizumab pegol in the management of rectovaginal fistulae due to Crohn’s disease). A more relevant focused question here may be: What is the efficacy of antitumor necrosis factor agents in treating adults with moderate–severe luminal Crohn’s disease.

Developing a systematic review protocol. Once a focused question is developed, it should be translated into a well-defined, systematic review question based on PICO criteria: Patients, Interventions, Comparators, and Outcomes (to be considered for inclusion in the review). For example, a systematic review question could be “In adults with moderate–severe luminal Crohn’s disease, are antitumor necrosis factor agents effective in inducing and maintaining remission, as compared with placebo or non-tumor necrosis factor-based therapies?”. Subsequently, a detailed systematic review protocol should be drafted, which includes the following: (a) inclusion and exclusion criteria (based on PICO, what types of study design, study durations, co-interventions, etc.), (b) a reproducible search strategy, developed in conjunction with an experienced medical librarian, detailing databases to be searched, including time frames, as well as a search terms with a combination of key words and medical subject headings, (c) *a priori*-planned hypothesis with subgroup and sensitivity analyses to identify potential sources of heterogeneity, (d) data abstraction elements, divided into study, patient, intervention/comparator, and outcome characteristics, (e) approach to risk of bias assessment, and (f) statistical approach. Ideally, these protocols should undergo either peer-review or be made publically available on open-access platforms (for example, PROSPERO, an international prospective register of systematic reviews).

Conducting a literature search. A systematic literature search of multiple databases, through controlled vocabulary terms and concepts, is an integral part of systematic reviews and avoids biases inherent to narrative reviews. Different databases have intrinsic differences.[9] For example, MEDLINE produced by the US National Library of Medicine focuses on articles in peer-reviewed journals of biomedicine and health, whereas Embase includes broader coverage of drugs and pharmacology and conference abstracts. CINAHL is an excellent source for research of nursing, allied health, or interprofessional areas. PsycINFO is the primary database for literature in psychology, psychiatry, counseling, addiction, and behavior. Regional and national databases may be important in searching for certain topics (e.g., searching LILACS, a database from Latin America and the Caribbean, when evaluating a tropical disease; searching the Chinese Biomedical Literature Database when evaluating a complementary medicine topic). Extensive literature reviews can be difficult to perform for most clinicians, and engaging medical librarians in the systematic review process significantly improves the quality of the search.[9] It is important that the initial search approach be sensitive (to avoid missing any potentially relevant article), rather than specific. Besides searching articles published in peer-reviewed literature, clinical trial registries, and relevant conference proceedings, as well as searching gray literature, is critical to minimize the "file-drawer" problem (negative studies may not be published in full).

Screen titles and abstracts and identify full texts for inclusion. Once the systematic literature review is completed, two sets of investigators review the title and abstracts, independently, to identify potential articles of interest (based on prespecified inclusion/exclusion criteria in the protocol). Full texts of articles are then reviewed to confirm inclusion. It is imperative to perform this step in duplication and, independently, to avoid bias in study selection, and degree of chance-adjusted agreement (kappa coefficient) between investigators should be noted. Conflicts in study identification should be resolved in conjunction with a third investigator. A detailed assessment of why studies are excluded, particularly when selecting after full-text review, is helpful.

Data abstraction. After study identification, two sets of investigators independently abstract key study, patient, intervention, and outcome variables in predesigned and piloted data abstraction forms. Independent data abstraction minimizes risk of errors. This could be performed on paper or electronic forms, such as Microsoft Excel or Word, or through data systems such as DistillerSR, Systematic Review Data Repository, etc. Details of different data collection tools are summarized elsewhere.[10] Sometimes, especially for larger reviews, a single reviewer may perform the primary data abstraction, and the secondary reviewer may independently abstract key data elements (particularly those that go into analysis) and confirm abstraction of random data elements across multiple studies. Occasionally, individual studies may not report all pertinent outcomes or after adjustment for confounders, due to space constraints in journal or other factors (even though it is likely that the data are available/analyzed); in these instances, earnest attempts should be made to contact study authors to obtain pertinent data to facilitate appropriate synthesis.

Risk of bias assessment. A key purpose of systematic reviews is to critically and objectively appraise risk of bias in relevant literature (also known as quality of individual studies) to facilitate appropriate interpretation of the body of evidence. This may be carried out using different proposed risk of bias tools for different study designs and are summarized elsewhere;[11] this assessment should also be performed in duplicate and independently. It is important to recognize that not all elements in risk of bias assessment tools are weighted equally, and the relative importance of different elements may vary depending on clinical question. For example, a seemingly "high-quality" population-based cohort study may fail to adjust for key confounders, which can bias evidence. In these instances, a cutoff to define "high" vs "low" quality studies may be misrepresentative but rather a qualitative critical assessment is more important.

Quantitative synthesis or meta-analysis. Although a decision on whether or not to perform quantitative synthesis is generally made in the review protocol, this may be revisited after identifying studies, appraising risk of bias, and abstracting data. One important reason not to perform a meta-analysis would be considerable conceptual heterogeneity in studies (systematic differences in study design, patient populations, interventions, or co-interventions, including study duration, drug dosing, and outcome assessment) where the investigators deem the studies are systematically different from one another such that quantitative synthesis would not be generalizable and applicable to clinical practice; this decisions should not be based on the presence or absence of statistical heterogeneity (assessed after performing quantitative synthesis). Once meta-analyses is being performed, important aspects include:

Generate summary estimates and confidence intervals. Meta-analysis allows estimate of a summary estimate of effect (reported as odds ratio, relative risk, or hazard ratio for comparative studies, and pooled proportions, prevalence, or incidence for single-group studies) and corresponding confidence intervals. Traditional DerSimonian–Laird statistical models in meta-analyses include fixed-effects and random-effects models.[12] The latter is generally preferred as it is conservative and factors both within- and between-study heterogeneity, and often results in a wider confidence intervals; the former may be used if the included studies are nearly identical and the number of studies is small (<5).[13] In the absence of heterogeneity, results of random- and fixed-effects model are identical. Again, decision on model to be used is not data-driven, but more concept-driven, and are made *a priori*.

Identification of sources of heterogeneity. With quantitative synthesis, a statistical measure of inconsistency, such as inconsistency index (I^2), should be reported.[14] This measures

Table 2 Factors to consider when interpreting credibility of claims of significance from subgroup analyses

	Criteria to consider claims on subgroup analyses	Interpretation
1	Can chance explain the subgroup differences?	Instead of focusing on each subgroup in isolation, compare summary point estimate and confidence intervals across subgroups—if point estimate is similar, and confidence intervals are overlapping, and statistical test of interaction. For example, if subgroup 1 has effect estimate (RR) 0.78 with 95% CI 0.40–1.05, and subgroup 2 has effect estimate 0.75 with 95% CI 0.38–0.95, then the correct interpretation would be that there is NO difference in subgroups, rather than that effects are significant in subgroup 2 and not in subgroup 1; results may not be statistically significant in subgroup 1 due to small sample size or low event rate in subgroup 1, rather than true differences in efficacy of intervention in subgroups
2	Is the subgroup difference consistent across studies and suggested by comparisons within rather than between studies?	Findings from subgroup analyses are credible if observed in multiple individual studies, rather than just at summary level
3	Was the subgroup difference one of a small number of *a priori* hypotheses in which the direction was accurately prespecified?	Prespecified, hypotheses-driven subgroup analyses to explain heterogeneity across studies are more plausible, rather than *post hoc* assessment, which may be positive due to multiple statistical comparisons
4	Is there a strong preexisting biological rationale supporting the apparent subgroup effect?	Subgroup claims are more credible if supported by strong external, biological evidence from preclinical studies or studies of surrogate outcomes

CI, confidence interval; RR, relative risk.

what proportion of total variation across studies was due to heterogeneity rather than by chance; here a value of <30%, 30–60%, 61–75%, and >75% is suggestive of low, moderate, substantial, and considerable heterogeneity, respectively. *A priori*-hypothesized subgroup analyses can facilitate identification of sources of heterogeneity or inconsistency of studies; while this assessment is definitely important if heterogeneity is identified, this should also be performed even if there is no considerable heterogeneity, to present stability of association across subgroups. One common misunderstanding in performing and interpreting subgroup analyses is overinterpretation of findings, resulting in spurious claims. Key factors to consider in differentiating credible and less credible subgroup analyses claims are reported in Table 2.[15] Quite often, subgroup analyses may not adequately explain observed heterogeneity and warrants a critical and qualitative assessment why studies may be inconsistent; this should be duly acknowledged and confidence in effect estimates tempered when drawing conclusions on body of evidence.

Small study effects, including publication bias assessment. Besides systematic attempts at comprehensive literature review to minimize risk of missing studies, statistical assessment of study effects, wherein a few small studies (frequently with large effect estimates) may influence overall effect estimate, is generally recommended.[13] This may be through visual inspection of funnel plot analysis or quantitative tests, such as Egger's regression test.[16] Of note, these tests are often underpowered and may not yield valid results if the number of studies is small (<10) or if there is considerable statistical heterogeneity.

NEWER CONCEPTS IN SYSTEMATIC REVIEWS AND META-ANALYSIS

Network meta-analysis. There is general paucity of head-to-head trials of active interventions, comparing different pharmacological interventions, which can inform stakeholders regarding the comparative effectiveness of these interventions, an oft-faced clinical dilemma. Traditional, direct pairwise meta-analyses provide only partial information in this case, because they can only answer questions about pairs of treatments and, hence, do not optimally inform decision-making. To overcome limitations in this, network meta-analyses have recently gained prominence.[17,18] These can help assess comparative effectiveness of several interventions and synthesize evidence across a network of randomized controlled trials. This method involves the simultaneous analysis of direct evidence (from randomized controlled trials directly comparing treatments of interest) and indirect evidence (from randomized controlled trials comparing treatments of interest with a common comparator) to calculate a mixed-effect estimate as the weighted average of the two.[19] Such a technique can improve the precision of the estimate (compared with direct evidence alone) and also allows estimation of the comparative efficacy of two active treatments, even if no studies directly compare them. Bayesian network analysis combines likelihood with a prior probability distribution to estimate a posterior probability distribution. For example, through a Bayesian network of three agents A, B, and C, if we know the relationship between A and B and between B and C, we can infer the probabilistic relationship between A and C (Figure 1). This allows us to estimate comparative treatment effects of two agents that have not directly been compared against each other but each has been compared against a common comparator (for example, a placebo). When considering performing a network meta-analysis, it is critical that included trials be conceptually similar in terms of key factors that determine treatment efficacy, including patients (similar disease characteristics and severity, prior failure of therapies), included interventions (standard dose and schedule), co-interventions (which can influence treatment efficacy), and outcome assessment

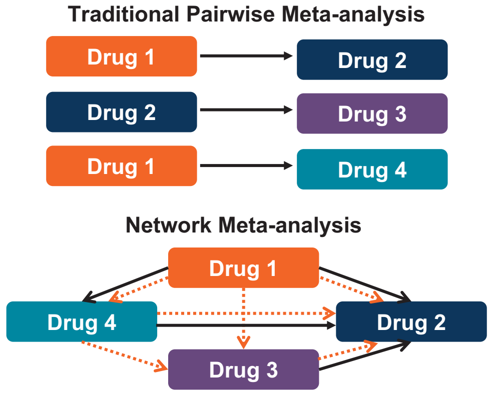

Figure 1 Differences between traditional meta-analyses and network meta-analyses. In traditional pairwise meta-analysis, only head-to-head direct comparisons can be analyzed. In contrast, network meta-analyses involve the simultaneous analysis of direct evidence (from randomized controlled trials (RCTs) directly comparing treatments of interest, indicated by solid arrows) and indirect evidence (from RCTs comparing treatments of interest with a common comparator, indicated by dotted arrows) to calculate a mixed-effect estimate as the weighted average of the two.

(similar reporting indices and definitions for outcome, assessed in standard manner).

Assessing quality of evidence. Beyond understanding the credibility of systematic reviews, it is imperative that investigators performing these, and clinicians reading these, critically appraise and interpret the quality of the body of evidence or the confidence in the summary effect estimate. Several systems are available, including the Grading of Recommendations Assessment, Development and Evaluation (GRADE), systems from the American Heart Association, US Preventive Services Task Force, and the Oxford Center for Evidence-based Medicine, and are summarized elsewhere.[7] GRADE categorizes evidence as high, moderate, low, or very-low quality.[20] The lower the confidence, the more likely the underlying true effect is substantially different from the observed estimate of effect and more likely is future research to demonstrate different estimates. In this approach, direct evidence from randomized controlled trials starts at high quality and confidence can be rated down if there is high risk of bias in the evidence, inconsistency (or heterogeneity) across studies, indirectness in evidence (i.e., results of meta-analysis are applied to same population from they are derived), imprecision (in case of small studies with low number of events, increasing fragility of summary effect estimate), and/or publication bias to levels of moderate, low, and very-low quality.

In summary, well-conducted systematic reviews and meta-analyses on focused pertinent clinical questions, with critical appraisal of body of evidences, can be are very useful in clinical practice and can inform clinical guidelines.

CONFLICT OF INTEREST

Guarantor of the article: Siddharth Singh, MD, MS.
Financial support: Dr Singh is supported by the NIH/NLM training grant T15LM011271, the American College of Gastroenterology Junior Faculty Development Award, and Crohn's and Colitis Foundation of American Career Development Award. Dr Singh has received research grants from Pfizer.
Potential competing interests: None.

1. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005; **2**: e124.
2. Fanelli D, Ioannidis JP. US. studies may overestimate effect sizes in softer research. *Proc Natl Acad Sci USA* 2013; **110**: 15031–15036.
3. Pfeiffer T, Bertram L, Ioannidis JP. Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLoS ONE* 2011; **6**: e18362.
4. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005; **294**: 218–228.
5. Lehrer J. The truth wears off. The New Yorker, 13 December 2010 (Accessed 1 March 2017).
6. Murad MH, Montori VM. Synthesizing evidence: shifting the focus from individual studies to the body of evidence. *JAMA* 2013; **309**: 2217–2218.
7. Murad MH, Montori VM, Ioannidis JP *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**: 171–179.
8. Moher D, Liberati A, Tetzlaff J *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; **151**: 264–269, W64.
9. Rethlefsen ML, Murad MH, Livingston EH. Engaging medical librarians to improve the quality of review articles. *JAMA* 2014; **312**: 999–1000.
10. Li T, Vedula SS, Hadar N *et al.* Innovations in data collection, management, and archiving for systematic reviews. *Ann Intern Med* 2015; **162**: 287–294.
11. Viswanathan M, Ansari MT, Berkman ND *et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions.* In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality: Rockville, MD, USA, 2008, pp 193–221.
12. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–188.
13. Higgins JPT, Altman DG, Sterne JAC (eds). Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration: New York City, NY, 2011.
14. Higgins JP, Thompson SG, Deeks JJ *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–560.
15. Sun X, Ioannidis JP, Agoritsas T *et al.* How to use a subgroup analysis: users' guide to the medical literature. *JAMA* 2014; **311**: 405–411.
16. Egger M, Davey Smith G, Schneider M *et al.* Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; **315**: 629–634.
17. Cipriani A, Higgins JP, Geddes JR *et al.* Conceptual and technical challenges in network meta-analysis. *Ann Intern Med* 2013; **159**: 130–137.
18. Mills EJ, Ioannidis JP, Thorlund K *et al.* How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA* 2012; **308**: 1246–1253.
19. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004; **23**: 3105–3124.
20. Guyatt G, Oxman AD, Sultan S *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**: 151–157.