Systematic review and meta-analysis methodology

Mark Crowther,[1] Wendy Lim,[2] and Mark A. Crowther[2]

[1]Department of Haematology, Worcestershire Royal Hospital, Worcester, United Kingdom; and [2]Department of Medicine, Division of Hematology-Thromboembolism, McMaster University and St. Joseph's Hospital, Hamilton, ON

Systematic reviews and meta-analyses are being increasingly used to summarize medical literature and identify areas in which research is needed. Systematic reviews limit bias with the use of a reproducible scientific process to search the literature and evaluate the quality of the individual studies. If possible the results are statistically combined into a meta-analysis in which the data are weighted and pooled to produce an estimate of effect. This article aims to provide the reader with a practical overview of systematic review and meta-analysis methodology, with a focus on the process of performing a review and the related issues at each step. (*Blood*. 2010;116(17): 3140-3146)

Introduction

The average hematologist is faced with increasingly large amounts of new information about hematologic disease. This ranges from the latest findings of complex molecular studies to results from randomized controlled trials (RCTs) to case reports of possible therapies for very rare conditions. With this vast amount of information being produced in published journals, presentations at conferences, and now increasingly online, it is virtually impossible for hematologists to keep up to date without many hours being spent searching and reading articles. For example, a search for 'deep vein thrombosis' in PubMed produced 55 568 possible articles, with 831 published in 2010 alone (search performed May 11, 2010). Review articles traditionally provide an overview of a topic and summarize the latest evidence, thus reducing the time clinicians would need to spend performing literature searches and interpreting the primary data. These review articles, known as narrative reviews, typically address a broad number of issues related to a topic.[1] Narrative reviews do not describe the process of searching the literature, article selection, or study quality assessment. The data are usually summarized but not statistically combined (qualitative summary), and key studies are highlighted. The inferences made from narrative reviews may be, but are not necessarily, evidence based. Narrative reviews are useful for obtaining a broad overview of a topic, usually from acknowledged experts. However, narrative reviews are susceptible to bias if a comprehensive literature search is not performed, or if only selected data are presented which conveys the author's views on a particular topic.[2]

Systemic reviews aim to reduce bias with the use of explicit methods to perform a comprehensive literature search and critical appraisal of the individual studies. Thus, in contrast to narrative reviews, systematic reviews pose a defined clinical question. The process of performing the literature search and the specific inclusion and exclusion criteria used for study selection are described. The quality of the included studies is formally appraised. The data are summarized, and, if the data are statistically combined (quantitative summary), the systematic review is referred to as a meta-analysis. The inferences made from systematic reviews are usually evidence based.

Furthermore, systematic reviews also attempt to identify if certain subtypes of evidence (eg, small negative studies) are absent from the literature; this so-called "publication bias" is an important cause of incorrect conclusions in narrative reviews.[3] Systematic reviews frequently, but not necessarily, use statistical methods, meta-analysis, to combine the data from the literature search to produce a single estimate of effect.[4]

In view of the increasing number of systematic reviews published, we feel it is important to discuss the methodology of the systematic review to allow readers to better appreciate and critically appraise systematic reviews that may be relevant for their practice. The findings of systematic reviews can be included in the introduction of scientific papers and are increasingly performed for grant applications to summarize what is known about a topic and highlight areas in which research is needed.

Having the knowledge to appraise a systematic review is an important skill, because systematic reviews are considered to be the study design with the highest level of study quality. Although many studies are labeled as systematic reviews, this does not necessarily indicate that the study itself is of high quality because any group of studies can be subject to a systematic review, and data can almost always be combined in a meta-analysis. The important issue is identifying if the systematic review was conducted in a manner that is replicable and free of bias, and if a meta-analysis was performed whether the data were appropriately combined. An evaluation of the quality of reviews (as measured by specific published criteria) published in 1996 in 6 core general medicine journals found that only 1% of review articles met all the recommended methodologic criteria.[5]

The objective of this article is to provide a practical approach to preparing and critically appraising a systematic review. Further guidance can be obtained from the Cochrane Collaboration's Web site,[6] and recommendations for reporting of systematic reviews are outlined by the Quality of Reporting of Meta-analyses (QUORUM) group (for randomized trials[7]), and the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group (for observational studies[8]). A modified version of the Quality of Reporting of Meta-analyses statement is presented in Table 1.

Submitted May 23, 2010; accepted July 1, 2010. Prepublished online as *Blood* First Edition paper, July 23, 2010; DOI 10.1182/blood-2010-05-280883.

Table 1. The QUORUM statement on how to report a systematic review

Heading/subheading	Descriptor
Title	Can clearly determine that report is a systematic review.
Abstract	Uses a structured format.
Objectives	Describes clinical question explicitly.
Data sources	Lists the databases used and other sources of data.
Review methods	Describes the process of how data was selection, quality assessment, data extraction, and any meta-analysis performed.
Results	Describes included and excluded studies and the results of any meta-analysis.
Conclusion	Describes the main results.
Introduction	Discusses the clinical problem, why the intervention may work, and the reasons for performing the review.
Methods	
Searching	Describes the data sources (eg, databases, handsearching, registers, researchers) and any search exclusions (date, language).
Selection	Inclusion and exclusion criteria outlined.
Validity assessment	Describes how any quality assessment was performed.
Data abstraction	Discusses how data was extracted from studies.
Quantitative data synthesis	Information on how data was combined (meta-analysis), including statistical methods used, measures of effect, and any sensitivity and subgroup analysis performed. Also explains which tests were performed, looking for heterogeneity and publication bias.
Results	
Trial flow	Provides a figure showing the number of studies screened, included, and excluded at each step.
Study characteristics	Each trial is described briefly, including participant demographics, number of participants, intervention, and follow-up.
Quantitative data synthesis	Presents simple summary results for individual studies and any meta-analysis performed.
Discussion	Discusses the answer to the original question in the light of the best available evidence and any possible biases. Also suggests future research.

Modified from Moher et al.[7]

The clinical question

What is the clinical question that needs to be answered? A careful articulation of the question is critical, because it provides the scope of the review by defining the type of patients, intervention, comparator, and outcomes evaluated in the review.[9] The nature of the question dictates study eligibility; hence, the more specific the question, the more focused the literature search albeit at the expense of decreased generalizability of the results. Thus, a systematic review on the use of colony-stimulating factors in patients with hematologic malignancies will be a far greater undertaking than a systematic review of colony-stimulating factors in preventing chemotherapy-induced febrile neutropenia in children with acute lymphoblastic leukemia.[10] When reading a systematic review one must always ascertain that the investigators are answering the question originally posed. A major cause of bias in a systematic review is answering a different question to that originally asked.

Search strategy

The completeness of the search strategy will determine the comprehensiveness of the review. The more exhaustive the search the greater the effort required to produce the systematic review, but the resulting review is generally of higher quality. The development of an inclusive search strategy requires expertise, and, unless the investigator is skilled in literature searches, the help of an experienced librarian is invaluable and strongly recommended. It is recommended that the search be performed in duplicate, because one person, especially if he or she is screening thousands of studies, may miss relevant studies. The literature search usually involves searching the following source.

Electronic databases

Many readers may remember the published *Index Medicus* in which journal articles were indexed based on topic. This has since been replaced by several electronic, Web-based, searchable databases. By entering a search strategy (usually according to Boolean language [OR, AND, NOT]) the databases provide a list of articles that meet the search criteria. The database to be used depends on what area of medicine the search is to be performed in. Examples of commonly used databases are shown in Table 2. The most commonly used databases include PubMed, MEDLINE, Embase, and the Cochrane library. MEDLINE is the largest component of PubMed, which is a free online database of biomedical journal citations and abstracts created by the US National Library of Medicine.

Conference abstracts

Many papers are presented at conferences before publication. It can take years for the content of these abstracts to be published. Conference abstracts can be searched and evidence can be extracted before full publication. The abstracts themselves may provide sufficient data to be included in the systematic review or, if a significant publication is anticipated, may warrant contacting the abstract author to obtain information. There are advantages and disadvantages to including conference abstracts. Studies that show inconclusive or negative results for an intervention are less likely to be published in journals but may be published in abstract form. Data from abstracts, reports, or other documents that are not distributed or indexed by commercial publishers (and which may be difficult to locate) are known as "gray literature."[11] Inclusion of abstracts and other gray literature potentially reduces the effect of publication bias. However, abstract results often differ significantly from the final publication, and abstracts have not generally undergone the rigorous peer review process required for most journal articles. This increases the likelihood that bias will influence the results of the systematic review.

Handsearching

The introduction and discussion section of relevant studies may provide additional references on a subject that may have been

Table 2. Examples of some commonly used electronic databases

Database	Contents
AIDS and Cancer Research (CSA)	Scientific literature related to AIDS, immunology, virology, and cancer genetics.
BioMed Central	All BioMed Central full-text journals.
CAB Abstracts	Research and development literature in the fields of agriculture, forestry, human and animal health, nutrition, and the management and conservation of natural resources.
Cochrane Library	Collection of randomized controlled trials and systematic reviews.
Cumulative Index to Nursing and Allied Health (CINAHL)	Journals, books, conference proceedings, pamphlets, and educational software of nursing literature.
Evidence-based medicine (EBM) reviews	Searches 4 evidence-based databases: ACP Journal Club (ACP), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), and Database of Abstracts of Reviews of Effectiveness (DARE).
Excerpta Medica (EMBASE)	Database of 3500 wide-ranging medical journals.
Health Services/Technology Assessment Text (HSTAT)	Evidence reports, treatment guidelines and protocols, and health technology assessments.
History of Science, Technology, and Medicine (OCLC)	Journal articles, conference abstracts, books, and dissertations on the history of science of medicine.
MEDLINE	Contents of 3900 health journals and conferences.
Social services abstracts (CSA)	Database of 1400 publications that focus on social welfare, social work, and human services.

missed by the search strategy. It is recommended that authors manually search the reference lists of found studies as a final check that no studies have been missed. One can also manually search journals in which studies on the subject of the review are likely to be published.

Contacting investigators

Writing to investigators active in the area may provide results of studies yet to be presented or published, but care must be taken with this information because it has not undergone any review process. Furthermore, most investigators will be hesitant to provide unpublished information because its inclusion in a systematic review may hamper subsequent publication. Perhaps the greatest utility of inquiring with investigators is gaining knowledge of studies about to be published, and when delaying the systematic review will allow inclusion of these articles and thus make the review more timely. Investigators may also be contacted if clarification of published information is required.

Internet

Apart from the searchable databases discussed earlier, there are other useful resources online. These include registers of clinical trials (eg, www.clinicaltrials.gov), data clearinghouses (eg, http://www.guideline.gov/), agencies charged with improving the quality of health care (http://www.ahrq.gov/), information on specific researchers from academic Web sites, university theses, and product information from drug companies. Searching the Internet with the use of search engines such as Google provides a user-friendly method to obtain information, but it is not recommended for systematic reviews because the accuracy of the information on the Internet is not ensured. Internet searches are notoriously nonspecific, and much time may be spent without much gain; this highlights the advantages of seeking input from a professional librarian before initiating a search.

Several decisions need to be made with regard to the search strategy. Investigators must decide what databases are to be searched, the level of detail in the search strategy, and whether the search should be done in duplicate. The search will produce a large number of possible studies, many of which can be excluded on the basis of their title and abstract. However, more detailed review of individual studies is required for those studies passing the initial screen.

When appraising a review, the reader should assess the completeness of the literature search. All relevant databases should be searched, and the search terms should be scrutinized for alternate terms or alternate spellings. For example, searching 'hemolytic anemia' in PubMed yields 58 560 citations, whereas 'haemolytic anemia' yields 3848 citations. The reader should also assess if the search strategy could have excluded relevant studies by being too specific. For example, when determining the effect of intravenous immune globulin (IVIg) in immune thrombocytopenia, a search strategy might be 'IVIg AND immune thrombocytopenia,' but this may miss studies that looked at patients with immune thrombocytopenia who were treated with steroids and given IVIg if they did not respond. If these studies were important to the question being asked, a more general search strategy might be considered, such as 'immune thrombocytopenia treatment.'

Study selection

It is best to establish, a priori, inclusion and exclusion criteria for accepting studies. These criteria should be explicit, and the most rigorous reviews should record the specific reasons for including or excluding all studies identified in the literature search. Specific recording for each study not only reduces the risk of bias but also allows rapid reassessment should the rationale for exclusion of one or more studies be called into question. Selecting studies in duplicate can help ensure that the correct studies are included and relevant studies are not missed. Agreement statistics can be calculated on the selection process, most commonly using the κ statistic.[12]

Deciding which studies to include or exclude in the review is very important. Inclusion or exclusion is usually based on the following different reasons.

Study design

The quality of the systematic review or meta-analysis is based in part on the quality of the included studies. RCTs have a lower potential for bias compared with observational studies. If there are available RCTs in the area of the review, the included studies may be limited to RCTs. Generally, RCTs (studies in which participants are randomly assigned to an intervention) are intrinsically of better quality than nonrandomized studies (in which participants are given an intervention then compared with another group that is similar but did not receive the intervention) that in turn are better than case series or case reports. The randomization process should equally distribute measurable and unmeasurable confounding factors between the 2 groups. As a result, differences observed should

Table 3. An example of inclusion and exclusion criteria

	Inclusions	Exclusions
Participants	Aged older than 16 years	Diseases with disease-specific immunodeficiencies (eg, CLL, myeloma, HIV-associated lymphoma)
	Lymphoma confirmed on biopsy (based on predefined classifications)	
Intervention	G-CSF or GM-CSF at a dose of at least 1 μg/kg/d given as primary prophylaxis for nonmyeloablative neutropenia, given within 72 hours after chemotherapy	Sequential administration of G-CSF or GM-CSF
	First- or second-line chemotherapy	Secondary prophylaxis
	Both groups to receive identical supportive care	Use during established neutropenia
		Stem cell transplantation
Types of studies	Randomized controlled studies comparing G-CSF or GM-CSF with placebo or no prophylaxis	Crossover studies, quasi-randomized studies (allocation by nonrandom methods, eg, DoB, name) or nonrandomized studies
	Abstracts or unpublished work with sufficient information.	Abstracts or unpublished work with insufficient information
Outcomes to be measured (requires at least 1 for inclusion)	Overall survival	
	Freedom from treatment failure	
	Quality of life	
	Risk and duration of neutropenia	
	Risk and duration of febrile neutropenia	
	Infection	
	Mortality	
	Etc	

An example of inclusion and exclusion criteria for studies to be included in a systematic review of colony-stimulating factor use for the prevention of adverse effects in the treatment of lymphoma.[13]

CLL indicates chronic lymphoblastic leukemia leukemia; G-CSF, granulocyte colony-stimulating factor; GM-CSF, granulocyte macrophage; and DoB, date of birth.

be due to the intervention rather than their occurring as a result of the effect of differences between those patients receiving the experimental, and those receiving the control, intervention. Given the reduced likelihood for bias, many systematic reviews only include RCTs; nonrandomized data are only included if randomized data are not available. It is important to note that even if the systematic review is based on randomized data, this does not ensure that the review itself is of high quality or that definitive conclusions can be made.

Language

Limiting studies by language will reduce the number of studies needed to review, especially if there is difficulty in translating a study. This may be acceptable for many reviews, but in some areas there may be many important studies published in other languages. Consequently, excluding studies on the basis of language must be done with care. For example, Chagas disease is endemic in Latin America, and a systematic review of transfusion-transmitted Chagas disease limited to English-only publications will exclude potentially important studies.

Date of publication

Limiting studies by date can be done if data does not exist before a specific date. For example, imatinib mesylate (Gleevec) for treatment of chronic myeloid leukemia was developed in the 1990s with phase 1 clinical trial data emerging by the end of the decade. Hence, a systematic review that involves imatinib would not require a literature search earlier than 1990.

Duplicate data

Some studies publish interim data or use the same patient cohorts in multiple publications. Excluding duplicate studies will eliminate overrepresentation of that particular data in the systematic review.

We would encourage readers to access the Cochrane Library's free-to-access database of systematic reviews of hematologic malignancies[13] for detailed examples of study inclusion and exclusion criteria. A summary of inclusion and exclusion criteria for granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma[14] can be found in Table 3.

Assessing the quality of studies

The quality of the studies included in the systematic review determines the certainty with which conclusions can be drawn, based on the summation of the evidence. Consequently, once all the relevant studies have been identified, the studies should undergo a quality assessment. This is particularly important if there is contradictory evidence. As with study selection, quality assessment performed in duplicate can help to minimize subjectivity in the assessment.

Various tools are designed for performing study quality assessment. The Jadad score is frequently used for quality assessment of RCTs,[15] and the Newcastle-Ottawa score is used for nonrandomized studies.[16] A modified example of the Jadad score is seen in Table 4. The important features in a quality assessment of RCTs include the following:

- The participants are not highly selected and are similar to those found in normal clinical practice (content validity).
- Neither the participants nor the researchers are able to tell how patients will be allocated before random assignment (allocation concealment).[17]
- Participants are followed up for an appropriate length of time, depending on the outcome assessed.
- Follow-up should be complete, with as few participants as possible being lost to follow-up. The reasons accounting for why

Table 4. The modified Jadad scoring system for randomized controlled trials[15]

Questions	Score
1. Was the study described as randomized? If yes, score 1 point.	
2. If yes to question 1, was an appropriate randomization sequence described and used (eg, table of random numbers, computer generated, etc.)? If yes, score 1 point.	
3. If yes to question 1, was an inappropriate method to generate the sequence of randomization used (patients were allocated alternately, or according to date of birth, hospital number, etc.)? If yes, subtract 1 point.	
4. Was the study described as double blinded? If yes, score 1 point.	
5. If yes to question 4, was an appropriate method of blinding used (eg, identical placebo, active placebo, dummy, etc.)? If yes, score 1 point.	
6. If yes to question 4, was an inappropriate method for blinding used (eg, comparison of tablet vs injection with no double dummy)? If yes, subtract 1 point.	
7. Were the withdrawals and dropouts described? If yes, score 1 point.	

patients dropped out or were lost should be provided to assess if these losses were due, in whole or in part, to side effects from the treatment.

- As many people as is feasible who are involved in the study are masked to the treatment received; ideally, participants, care providers, data collectors, and outcome adjudicators should be masked. The statistician can also be masked to the specific intervention.
- The results of the study should be analyzed as intention to treat (all patients who underwent allocation are analyzed regardless of how long they stayed in the study; this provides the best "real world" estimate of the effect) and per protocol (only patients who remained within the protocol for a predetermined period are analyzed; this gives the best safety data).

When appraising a review, the reader needs to assess if an appropriate quality assessment tool has been used, whether the quality assessment was done in duplicate, and, if so, whether there was agreement between the investigators. Again, for useful examples in the setting of hematologic disease we encourage review of the Cochrane Library. For example, a systematic review on the use of immunoglobulin replacement in hematologic malignancies and hematopoietic stem cell transplantation[18] assesses the quality of the studies on the basis of allocation concealment, allocation generation (randomization procedure), and masking. In the sensitivity analysis that compared studies fulfilling and not fulfilling these criteria they showed no difference in the result.

Data extraction

The data from the studies can then be extracted, usually onto prepared data case report forms. The data to be extracted should be carefully considered before the start of the review to avoid having to re-extract data that were missed on the initial data collection. Data extraction should ideally be done in duplicate to allow identification of transcription errors and to minimize any subjectivity that may occur when interpreting data that are presented in a different format than that required on the case report form.

Combining the data (meta-analysis)

If suitable, data from several studies can be statistically combined to give an overall result. Because this overall result reflects data from a larger number of participants than in the individual studies, the results are less likely to be affected by a type 2 error (failing to detect a "real" difference that exists between the 2 groups). One of the main criticisms of meta-analysis is that studies that are quite different can have their results combined inappropriately, and the result is not an accurate reflection of the "true" value. Therefore, before embarking on a meta-analysis, one must consider whether the difference between the studies (heterogeneity) precludes pooling of the data. Only studies with similar interventions, patients, and measures of outcomes should be combined. For example, 3 studies that all evaluate the efficacy of a new iron-chelating agent in patients with transfusion-related iron overload may be inappropriately combined if one study compared the iron chelator with placebo and measured hepatic iron content by liver biopsy at 1 year, the second study compared the new iron chelator with a different iron chelator and performed liver biopsy at 6 months, and the third study was an observational study in which patients receiving the new iron chelator underwent cardiac magnetic resonance imaging at the start of treatment and again at 1 year. Although all 3 studies are evaluating the efficacy of a novel iron chelator, the studies differ importantly. Because they differ in their design, comparator, timing of outcome measurement, and the method of outcome ascertainment combining, these data are inappropriate.

The most commonly used meta-analysis software is RevMan,[19] available from the Cochrane Collaboration. Categorical data (eg, number of remissions) or continuous data (eg, time to relapse) is entered into the program, combined, and then visually presented in a Forest plot (explained in Figure 1). Various statistical methods are used for combining different types of data. The data from each individual study are weighted such that studies that have less variance (spread of data) or a larger sample size contribute more heavily to the overall estimate of effect. The common mathematical methods used to combine data include the Mantel-Haenszel method[20] and the Inverse Variance method.[21] The Mantel-Haenszel method is used for categorical data and results in a risk ratio or relative risk. The risk ratio expresses the chance that an event will occur if the patient received the intervention compared with if they received the control. For example, if a meta-analysis of studies that measure infections after giving prophylactic antibiotics to neutropenic patients showed a relative risk of 1.5 comparing no antibiotics with antibiotics, this means that taking antibiotics reduces the risk of infection by 1.5 times.

The Inverse Variance method is used for continuous data and results in a mean difference. The mean difference is the average difference that will be achieved by giving the patient the intervention rather than the control. If the same studies also measured length of stay and patients who had taken antibiotics had a mean difference of −1.3 days compared with patients not taking antibiotics, this would suggest that on average patients taking antibiotics would stay for 1.3 days less than patents who did not take antibiotics. The mean difference is used if the outcome that is measured is the same in all the studies, whereas the standard mean difference is used if the outcomes are measured slightly differently. For example, if studies of postthrombotic syndrome (PTS) severity used the Villalta PTS scale,[22] then the mean difference can be used for the meta-analysis. However, if some of the studies used the Villalta PTS scale and others used the Ginsberg clinical scale,[23] a standard mean difference should be used.

As discussed earlier, one of the main problems with meta-analysis is that the studies being combined are different, resulting in heterogeneity.[24] There will always be some heterogeneity

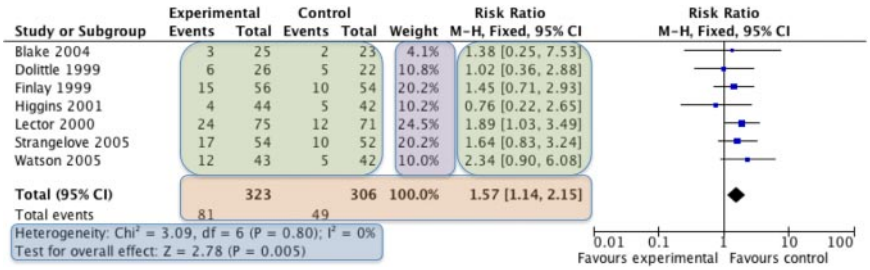

Figure 1. An example of a Forest plot. The names of the individual studies are on the left, the individual studies results are seen in the green boxes, and the overall combined result is seen in the orange box. The purple box shows the weighting given to each study, which is based on the number of participants (larger studies given more weight). The blue box displays the statistics for the meta-analysis, including whether the overall result is statistically significant (test for overall effect) and 2 measures of heterogeneity (χ^2 and I^2 tests). On the far right is the graphical representation of the results, known as the Forest plot. The studies are displayed horizontally, whereas the horizontal axis represents the magnitude of the difference between the intervention and control group. Each study is represented by a blue box and a black horizontal line. The blue box represents the result of the study, with the larger the box indicating the greater the weight of the study to the overall result. The black horizontal line represents the 95% confidence intervals for that study. If both the box and the horizontal line lie to the left of the vertical line, then that study shows that the intervention is statistically significantly better than the control, whereas, if the box and horizontal line all lie to the right of the vertical line, then the control is statistically significantly better. If the box or horizontal line cross the vertical line, then the individual study is not statistically significant. The overall result is represented by a diamond, with the size of the diamond being determined by the 95% confidence intervals for the overall combined result. If the diamond does not touch the vertical line, then the overall result is statistically significant, to the left the intervention is better than the control group and to the right indicates that the control group is better. If the diamond touches the line, then there is no statistical difference between the 2 groups.

between studies because of chance, but when performing meta-analysis this needs to be investigated to determine whether the data can be combined reliably. RevMan calculates 2 measures of heterogeneity, the χ^2 test and the I^2 test.[25] The χ^2 test determines whether there is greater spread of results between the studies than is due to chance (hence, heterogeneity is present) and a value less than 0.10 usually suggests this. The I^2 test tries to quantify any heterogeneity that may be present, a result greater than 40% usually suggests its presence, the higher the percentage the greater the heterogeneity. If heterogeneity is present, it should be investigated by removing studies or individual patients from the analysis and seeing if that removes the heterogeneity. Differences between the included patients in the individual studies may explain the heterogeneity (clinical heterogeneity). For example, the efficacy of a new treatment for multiple myeloma will depend on whether the patients' condition is newly diagnosed, previously treated, or after transplantation. Differences in drug dosing, route, and frequency of administration will also contribute to heterogeneity. Other contributors to heterogeneity may include the design of the study and how the study was funded (eg, commercial vs noncommercial sponsorship). If, after detailed investigation, there is no obvious cause for the heterogeneity, the data should be analyzed with a more conservative statistical method that will account for the heterogeneity. In reviews with significant heterogeneity, a more conservative overall result will be obtained if the analysis uses a random-effects model, compared with a fixed-effect model. A random-effects analysis makes the assumption that individual studies are estimating different treatment effects. These different effects are assumed to have a distribution with some central value and some variability. The random-effects meta-analysis attempts to account for this distribution of effects and provides a more conservative estimate of the effect. In contrast, a fixed-effect analysis assumes that a single common effect underlies every study included in the meta-analysis; thus, it assumes there is no statistical heterogeneity among the studies.

Other techniques used to account for heterogeneity include subgroup analyses and meta-regression. Subgroup analyses are meta-analyses on individual clinical subgroups that determine the specific effect for those patients. Common subgroups may be based on age, sex, race, drug dosage, or other factors. Ideally, subgroup analyses should be limited in number and should be specified a priori. Meta-regression is used to formally test whether there is evidence of different effects in different subgroups of studies.[26] This technique is not available in RevMan.

Sensitivity analyses can be performed to determine whether the results of the meta-analysis are robust. This involves the removal of studies that meet certain criteria (eg, poor quality, commercial sponsorship, conference abstract) to determine their effect on the overall result. For example, large studies will generally have a lower variance and will thus be more heavily weighted than small studies. However, because this weighting does not account for study quality, a poorly designed large study might be overrepresented in the analysis compared with a small well-performed study. A sensitivity analysis could be performed in which the large poorly designed study is removed and the meta-analysis is repeated with the remaining studies to assess if the overall effect estimate remains the same. For example, if a drug appears to have a positive effect on relapse-free survival, but this effect disappears when commercial studies are removed in the sensitivity analysis, the reader should be aware that the results and conclusions of such study could be biased.

Assessing for publication bias can be performed with funnel plots (Figure 2). Studies that are negative are less likely to be published, and their absence from the review is a potential source of bias.[11] In a funnel plot, the vertical axis measures the precision of the estimate of the treatment effect (eg, standard error of the log relative risk, sample size) and the horizontal axis measures the treatment effect (eg, relative risk). If there is no publication bias, all the studies should uniformly fall within the inverted V. If a section of the inverted V is devoid of studies, this indicates a publication bias (most often the failure of small negative studies to be published and thus included in the analysis). Analyses that fail to account for missing negative and smaller studies will tend to overestimate the treatment effect.

Making conclusions

When all the suitable studies have been collected, quality assessed, data extracted, and, if possible, meta-analysis performed, then conclusions need to be made. The authors must refer back to the original question and ask if there is enough evidence to conclusively answer the question and, if there is, how strong the

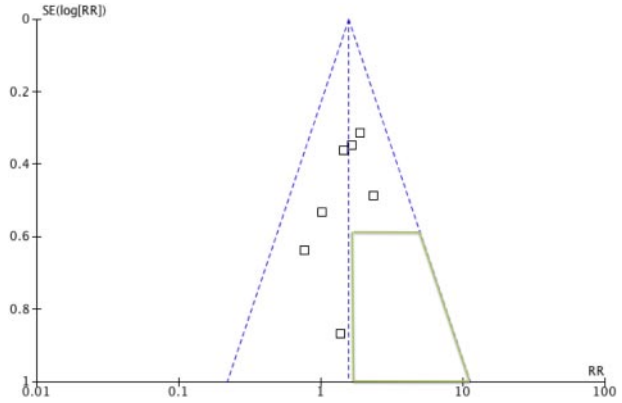

Figure 2. Funnel plot to assess for publication bias. In this plot the result of the individual study is plotted on the horizontal axis (in this case the risk ratio [RR]) against a measure of the precision of the data (either the spread of the data or the size of the study) on the vertical axis (this graph uses spread of data measured by the standard error of the log of the relative risk [SE(logRR)]), the smaller the spread of data or the greater the study size the further up the vertical axis. Individual studies are represented by the small squares. From the overall result of the meta-analysis the central estimate is plotted (the vertical dashed line), and the 95% confidence intervals are drawn (the diagonal dashed lines) to form the funnel or inverted V. The assumption is that the larger the study (or the study with a smaller spread of data) then the nearer to the true result it will be, meaning the spread about the overall result will be reduced as the study size increases, hence the funnel shape. If there is publication bias, then the studies will not be equally distributed within the inverted V. The usual sign of publication bias is the absence of studies in the green box that represents where small negative studies lie.

supporting evidence is. In evaluating a systematic review, the reader must decide if the authors have made an objective conclusion on the basis of the available evidence and not on personal opinion. The discussion should make reference to any sources of heterogeneity and whether there are subgroups in which the evidence is stronger than in others. The results of sensitivity analyses, if performed, may be discussed particularly if the results suggest the presence of bias in the overall results. At the end of the systematic review, it is possible that there is insufficient evidence to draw clear conclusions. In many situations, authors will conclude that further research is needed to provide stronger recommendations or give specific recommendations for clinically important subgroups.

Concluding remarks

We hope that you have found this article useful for future appraisal of systematic reviews and meta-analysis. The techniques described should not be used just for the production of "formal for publication" reviews but can be equally well applied to day-to-day analysis of clinical problems found in the consulting room. Systematic reviews, compared with primary research, requires relatively few resources, allowing clinicians not normally involved in research to produce high-quality, clinically relevant papers.

Authorship

Contribution: All authors contributed equally to the design, preparation, and editing of the document. M.A.C. was responsible for the final review and approval for submission.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Mark A. Crowther, Rm L301, St. Joseph's Hospital, 50 Charlton Ave East, Hamilton, ON, Canada, L8N 4A6; e-mail: crowthrm@mcmaster.ca.

References

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med*. 1997;126(5):376-380.
2. Schmidt LM, Gotzsche PC. Of mites and men: reference bias in narrative review articles: a systematic review. *J Fam Pract*. 2005;54(4):334-338.
3. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-872.
4. Hedges LK, Olkin I. *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press; 1986.
5. McAlister FA, Clark HD, van Walraven C, et al. The medical review article revisited: has science improved? *Ann Intern Med*. 1999;131(12):947-951.
6. The Cochrane Collaboration. http://www.cochrane.org. Accessed May 10, 2010.
7. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUOROM Group 1999. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. *Lancet*. 1999;354(9193):1896-1900.
8. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology. *JAMA*. 2000;283(15):2008-2012.
9. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak*. 2007; 7:16. doi: 10.1186/1472-6947-7-16. http://www.biomedcentral.com. Accessed May 10, 2010.
10. Sasse EC, Sasse AD, Brandalise SR, Clark OA, Richards S. Colony-stimulating factors for prevention of myelosuppressive therapy-induced febrile neutropenia in children with acute lymphoblastic leukaemia. *Cochrane Database Syst Rev*. 2005;(3):CD004139. http://www.thecochranelibrary.com. Accessed May 10, 2010.
11. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev*. 2007;(2):MR000010. http://www.thecochranelibrary.com. Accessed May 10, 2010.
12. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1986;76(5):378-382.
13. Cochrane Database of Reviews of Interventions in haematological malignancies. http://www.mrw.interscience.wiley.com/cochrane/cochrane_clsysrev_subjects_fs.html. Accessed May 24, 2010.
14. Bohlius J, Herbst C, Reiser M, Schwarzer G, Engert A. Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. *Cochrane Database Syst Rev*. 2008;(4):CD003189. http://www.thecochranelibrary.com. Accessed May 24, 2010.
15. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996; 17(1):1-12.
16. The Newcastle-Ottowa score for non-randomized studies. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed May 10, 2010.
17. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA*. 1994;272(2):125-28.
18. Raanani P, Gafter-Gvili A, Paul M, Ben-Bassat I, Leibovici L, Shpilberg O. Immunoglobulin prophylaxis in hematological malignancies and hematopoietic stem cell transplantation. *Cochrane Database Syst Rev*. 2008;(4):CD006501. http://www.thecochranelibrary.com. Accessed May 24, 2010.
19. Review Manager (RevMan) Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration; 2008.
20. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959(4):22:719-748.
21. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
22. Villalta S, Bagatella P, Piccioli A, et al Assessment of validity and reproducibility of a clinical scale for the post-thrombotic syndrome [abstract]. *Haemostasis*. 1994;24(suppl 1):Abstract 158a.
23. Ginsberg JS, Turkstra F, Buller HR, et al. Post-thrombotic syndrome after hip or knee arthroplasty: a cross-sectional study. *Arch Intern Med*. 2000;160:669-672.
24. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-5.
25. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557-560.
26. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-1573.