Systematic review and meta-analysis: techniques and a guide for the academic surgeon

Kevin Phan¹, David H. Tian¹, Christopher Cao¹, Deborah Black², Tristan D. Yan¹

¹The Collaborative Research (CORE) Group, Macquarie University, Sydney, Australia; ²Faculty of Health Sciences, University of Sydney, Sydney, Australia

Correspondence to: Tristan D. Yan. Professor of Cardiovascular and Thoracic Surgery, The Collaborative Research (CORE) Group, Macquarie University, Sydney, Australia. Email: tristanyan@annalscts.com.

With the rapidly growing literature across the surgical disciplines, there is a corresponding need to critically appraise and summarize the currently available evidence so they can be applied appropriately to patient care. The interpretation of systematic reviews is particularly challenging in cases where few robust clinical trials have been performed to address a particular question. However, risk of bias can be minimized and potentially useful conclusions can be drawn if strict review methodology is adhered to, including an exhaustive literature search, quality appraisal of primary studies, appropriate statistical methodology, assessment of confidence in estimates and risk of bias. Therefore, the following article aims to: (I) summarize the important features of a thorough and rigorous systematic review or meta-analysis for the surgical literature; (II) highlight several underused statistical approaches which may yield further interesting insights compared to conventional pair-wise data synthesis techniques; and (III) propose a guide for thorough analysis and presentation of results.

Keywords: Systematic review; meta-analysis; surgery; outcomes; forest plot; network meta-analysis; meta-regression

Submitted Feb 13, 2015. Accepted for publication Feb 14, 2015.

View this article at: http://dx.doi.org/10.3978/j.issn.2225-319X.2015.02.04

Systematic Review

Introduction

With the rapidly growing literature across the surgical disciplines, there is a corresponding need to critically appraise and summarize the currently available evidence to date so they can be applied appropriately to patient care (1,2). Evidence-based medicine is the explicit, conscientious and judicious use of the currently best available evidence from research to guide health care decisions. Clinical decisions should be based on the totality of the available evidence, rather than based on the results of any individual study or trial (3). In the evidence-based approach to surgery of the modern era, formal and comprehensive reviews of the literature, with or without additional statistical analysis, remains critically important to the practicing surgeon as a source of updated information on diagnoses, prognoses, and the effectiveness of healthcare interventions (4).

While the popularity of systematic review and meta-analyses is increasing, there are also limitations, which surgeons must keep in mind prior to applying their conclusions directly to patient care. The quality of systematic reviews and meta-analyses is a function of the quality of the primary studies available as well as the degree of rigor to which the systematic reviews have been performed (5). A systematic review or meta-analysis that has been poorly performed may produce misleading results and conclusions, using methods or statistical approaches which may lack credibility (6,7).

The conduct and interpretation of systematic reviews is particularly challenging in the case where few robust clinical trials have been performed to address a particular question. However, risk of bias can be minimized and potentially useful conclusions can be drawn if strict review methodology is adhered to, including an exhaustive literature search, quality appraisal of primary studies, appropriate statistical methodology, assessment of confidence in estimates and risk of bias.

Therefore, the following article aims to: (I) summarize...
to the important features of a thorough and rigorous systematic review or meta-analysis for the surgical literature; (II) highlight several underused statistical approaches which may yield further interesting insights compared to conventional pair-wise data synthesis techniques; and (III) propose a guide for thorough analysis and presentation of results.

### Framing the clinical question

Prior to the literature search, the review must ensure that the posed clinical question has a clear focus and is appropriate either for a systematic review or meta-analysis. It is important that the relevance and suitability of the question be carefully appraised so as to be better designed to improve current clinical knowledge, practice and guide policy and decisions.

A well-accepted methodology for this is the use of the PICO format, with a clearly defined study population (P), intervention studied (I), comparisons (C), and outcomes (O) (1). As an example, in a recent systematic review and meta-analysis on sutureless aortic valve replacement (8), the study population was defined as patients with aortic valve replacement requiring surgical intervention (P), the intervention was minimally invasive aortic valve replacement (I), the comparator was conventional aortic valve replacement (C), and outcomes include perioperative mortality and complications (O).

The challenge in developing a robust clinical question to set out to answer is defining the scope of patients and interventions. In order to define the scope of the question, the reviewers must have a comprehensive understanding of the existing literature, the potential gaps and uncertainties in the available evidence, and which gaps can be potentially addressed and answered via a systematic review or meta-analysis.

A scoping review (9) can be performed to explore the extent of the available evidence and to assist in determining the approach scope of the clinical question. Factors to take into consideration when developing the clinical question include the level of evidence and the design of available studies. For example, if there are an adequate number of randomized controlled trials (RCTs) available addressing the proposed clinical question, then the scope of the review should be limited to randomized studies to limit the effect of selection bias from non-randomized studies. In contrast, if there are few randomized trials available and that the evidence is mainly limited to observational studies, then the reviewers may consider broadening the scope of the review to include all observation and randomized trials. Subgroup analysis and sensitivity analysis can be performed as secondary analysis to elucidate any effect of their non-randomized design on the final effect sizes calculated (10,11). In the case of a meta-analysis, the question may be more narrow-focused, for example, with the inclusion of only trials that compare two particular interventions.

Similar logic also applies when considering the breadth or narrowness of scope for population, intervention, comparator and outcomes. For example, if the defined patient population is too broad, then any calculated estimate of effect that is generalized across a wide range of patients with varying risk factors could providing misleading trends and conclusions (12,13). For example, the durability of an implanted aortic valve would be different in young patients (<50 years) versus elderly patients (>70 years), based on their underlying comorbidities and operative risk, and thus pooling freedom from reoperation rates across all ages would providing an inaccurate indication of valve durability in different patient groups (14).

Without a clear question that is clinically relevant and has strictly defined population, intervention, comparator and outcome parameters, the systematic review performed risks being ambiguous, ill-structured, and heterogeneous with invalid interpretations of the results. We recommend the writing or publication of a research protocol prior to conducting the systematic review (15,16), with fully defined inclusion and exclusion criteria, participants, interventions, outcomes of interest and, strategy for statistical analysis.

### Literature search

Next, the authors should decide on the appropriate inclusion and exclusion criteria a priori, based on the clinical question to be answered. These criteria should be explicitly stated in the final manuscript, and may include the design of studies to be included (RCTs vs. any study type; comparative studies only), the study population in terms of gender, age group, disease, the language of the published studies, or the time period of publication (e.g., inclusion of studies published after year 2000). Exclusion may include, but not limited to, low levels of evidence such as abstract-only articles, conference articles, editorials and expert opinions, removal of studies with repeat follow-up reports of the same population, removal of studies which report fewer than 10 patients per arm, etc. The potential bias generated by these inclusion and exclusion criteria should be considered and
discussed along with the findings of the review.

To ensure a comprehensive and exhaustive literature search for appropriate primary studies, multiple databases should be systematically searched (17). Typical electronic databases include MEDLINE, EMBASE, PubMed, Cochrane Central Register of Controlled trials and ACP. Searching only a single electronic database is not recommended, as there is a high chance that relevant articles will be missed (18,19). It is important for the reviewers to use specific keywords and MeSH headings according to their clinical question posed, as well as Boolean operations such as “AND”, “OR”, and “NOT”. As per recommended PRISMA guidelines (20), an example of at least one search strategy should be provided in the final manuscript, either as one of the main text tables or as a supplementary table or appendix data. The literature search should be conducted by at least two reviewers independently. Any discrepancies in the final list of articles to be included should be discussed and resolved by consensus. Furthermore, additional references should be identified via searches of trial registries, reference lists of included studies, foreign language literature and contacting experts in the field (21-23). The overall search strategy should be presented as a PRISMA flow-chart in the final manuscript.

**Study quality appraisal**

In order to assess intra-study risk of bias, which may undermine the validity of the final results obtained, quality appraisal should be performed according to Cochrane and MOOSE guidelines. There are different checklists and tools available to assess intra-study risk of bias (24,25). The Cochrane Collaboration Review Manager software has an inbuilt tool for the assessment of RCTs (26), based on the following domains: sequence generation, allocation concealment, blinding of participants and outcomes assessment, incomplete outcome data, selective outcome reporting, and other sources of bias. RCTs are often regarded as high-quality “gold-standard” studies (27) that systematic reviews and meta-analyses should ideally include. However, different RCTs can have variations in methodological and reporting quality, and thus it is important to assess all RCTs for quality appraisal. High-quality RCTs should follow the reporting standards set out by the Consolidated Standards of Reporting Trials (CONSORT) statement, which includes a 22-item checklist and flow diagrams (28).

There are also multiple checklists available for the assessment of observational studies (29-31). The Dutch Cochrane Collaboration Group has developed one checklist and another commonly used checklist is STROBE (29). The key domains assessed by the MOOSE tool includes (31): (I) clear definition of study population; (II) clear definition of outcomes and outcomes assessment; (III) independent assessment of outcome parameters; (IV) sufficient follow-up; (V) no selective loss during follow-up; and (VI) important confounders and prognostic factors identified. It is highly recommended that these checklists be used to rigorously assess the quality of included studies. Completed checklists should be included as a table in the final manuscript, or as a supplementary table. The risk of bias assessment should be performed by at least two reviewers, and any discrepancies should be resolved by consensus. An example of such a risk assessment checklist is demonstrated in Table 1 from recent systematic reviews (8,32) in surgery. Studies which do not fulfil the a priori quality requirements should be carefully appraised to determine suitability of inclusion in further statistical analyses, such as via sensitivity analysis or cumulative meta-analysis (33-35) to detect heterogeneity or adjustment in effect size with time (34,36,37).

**Data extraction**

Data extraction should be performed according to a template with relevant demographic, operational parameters and outcomes that have been defined prior to review. Data extraction should be performed by at least two reviewers, and discrepancies resolved by consensus. This will reduce the risk of reviewer bias, error and subjectivity.

**Statistical approaches**

**Summary statistics**

The statistical techniques employed will depend on the type of review performed and available data. In the case of a systematic review of a single surgical intervention without comparison, it would be appropriate to report descriptive summary statistics in the form of mean, standard deviation, and range for parametric continuous measures. If the demographics of the study populations and inclusion/exclusion are relatively similar between the studies, this may warrant a meta-analysis of weighted proportions, using a random-effects model to pool one-arm cohort studies. An example of such analysis was performed to determine the weighted pooled paravalvular leak rate from sutureless aortic valve replacement (SU-AVR) (8) at 12-month follow-up (Figure 1).
Table 1 Example of an intra-study risk of bias assessment table, according to Dutch Cochrane Group and MOOSE guidelines. Each cell can be filled in with either yes/no/unclear. Study 1 and 5 would be deemed of higher quality as all criteria was rated as “yes” for being free from intra-study bias.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear definition of study population?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Clear definition of outcomes and outcome assessment?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Independent assessment of outcome parameters?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Unclear</td>
<td>Yes</td>
</tr>
<tr>
<td>Sufficient duration of follow-up?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>No selective loss during follow-up?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Important confounders and prognostic factors identified?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Studies</th>
<th>Estimate (95% C.I.) Ev/Tt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santarpino 2014</td>
<td>0.006 (0.000, 0.041) 0/78</td>
</tr>
<tr>
<td>Shrestha 2013</td>
<td>0.080 (0.028, 0.174) 4/50</td>
</tr>
<tr>
<td>Kocher 2013</td>
<td>0.007 (0.000, 0.029) 1/146</td>
</tr>
<tr>
<td>Glmanov 2013</td>
<td>0.058 (0.024, 0.105) 8/137</td>
</tr>
<tr>
<td>Echstaedt 2013</td>
<td>0.050 (0.017, 0.097) 6/120</td>
</tr>
<tr>
<td>Concire 2013</td>
<td>0.019 (0.000, 0.044) 1/97</td>
</tr>
<tr>
<td>Folguert 2012</td>
<td>0.010 (0.000, 0.029) 2/208</td>
</tr>
<tr>
<td>Martens 2011</td>
<td>0.007 (0.000, 0.030) 1/140</td>
</tr>
<tr>
<td>Flameng 2011</td>
<td>0.062 (0.001, 0.179) 2/32</td>
</tr>
<tr>
<td>Doss 2005</td>
<td>0.042 (0.000, 0.170) 1/24</td>
</tr>
<tr>
<td>Overall (P²=58%, P=0.012)</td>
<td>0.022 (0.008, 0.041) 26/1032</td>
</tr>
</tbody>
</table>

Figure 1 Meta-analysis of proportions of single-arm studies reporting paravalvular leak outcomes for SU-AVR, using a random-effects model. The weighted mean paravalvular leak rate here was 3.0%, with 95% confidence interval of 1.0-5.8%. Significant heterogeneity was demonstrated (I²=72%, P<0.001). SU-AVR, sutureless aortic valve replacement.

For a systematic review and meta-analysis of comparative studies, presentation of summary statistics in the form of Forest plots is appropriate (38). Forest plots involve a weighted compilation of all the effect sizes reported by each study, and also provide an indication of heterogeneity between studies. An example is shown in Figure 2, from a recent meta-analysis comparing ministernotomy versus minithoracotomy approaches for minimally invasive aortic valve replacement (39). For each study, the effect size is represented by a square and horizontal line, representing the point estimate and 95% confidence interval, respectively. The size of the square is proportional to the weight assigned to that particular study for the meta-analysis. The pooled effect size following meta-analysis is represented by the black diamond, the width of which is indicated of the overall 95% confidence interval. If this diamond lies totally to one side of the solid vertical line in the center, then the pooled point estimate indicates a significant difference in effect size between the two interventions (38).

Typical summary statistics used for point estimates include relative risk (RR) or odds ratio (OR) for dichotomous parameters, and weighted mean difference (WMD) for continuous data. The analysis can be performed using fixed-effect or random-effects models. In the prior, the true effect size is assumed to be similar among the included studies, whilst in the latter, the included studies represent a random sample of effect sizes. The random-effects model is most appropriate in the case where there is risk of heterogeneity in the effect size reported.

In order to assess whether effect sizes were consistent across the included studies, heterogeneity should be quantified (40). There are two commonly used tests for the assessment of heterogeneity. The Cochran Q test provides a yes vs. no outcome for whether there is significant
Another technique used to assess heterogeneity is meta-regression analysis. This statistical approaches determines whether there is a significant association between an independent variable in form of study or intervention characteristics (for example age, study time point, operation duration) versus the dependent variable, the outcome of interest. A regression model is constructed, and the P value and regression coefficient (r) can be used to assess the strength of this association. A significant relationship may indicate that the study variable may be a source of variability observed (8,43-45). For example in Figure 3, meta-regression analysis was used to show a significant negative correlation between midpoint of study period with rate of paravalvular leak for the Perceval S sutureless prosthesis (8). This suggests that the learning curve of SAVR may be a source of heterogeneity in the pooled results for paravalvular leak.

**Network meta-analysis**

Network meta-analysis may be appropriate for assessment multiple interventions (>2) for the same disease or outcome. Also known as multiple-treatment comparison, network meta-analyses aim to pool all available direct and indirect comparisons.
comparisons for multiple interventions to provide an overall comparison (46-48). In contrast to traditional pair-wise meta-analysis, the advantage of network meta-analysis is that using indirect evidence, all available data can be synthesized to provide comparative effect estimates between interventions where there may not have been direct head-to-head trials performed to date.

While the technical statistical details of network meta-analyses is beyond the scope of the present article, it is suffice to say that available direct evidence (A versus B) and indirect evidence (A versus C, C versus B) can be applied to a “Bayesian” statistical model where Monte Carlo simulations are run (49). The model will converge to the likelihood effect estimate and provide a modeled comparison between A versus B versus C. The underlying assumption of this approach is that the comparator group for the interventions (i.e., C) is similar amongst the indirect comparison trials (50).

Typical software packages for network meta-analyses include WinBUGS and GeMTC packages. The general work process for network meta-analysis is similar to traditional meta-analysis, involving: (I) data extraction from direct and indirect comparative studies; (II) importing the data into the software package e.g., WingBUGS or GeMTC; (III) running the Bayesian model and Monte Carlo simulations.

In the first systematic review to compare median sternotomy versus ministernotomy versus minithoracotomy for minimally invasive aortic valve replacement, a Bayesian network meta-analysis was performed based on direct and indirect evidence (39). This is particularly pertinent, given so few studies were available providing a head-to-head comparison between ministernotomy and minithoracotomy. As such, network meta-analysis was performed to produce an integrated effect size for ministernotomy vs. minithoracotomy, based on all available direct and indirect evidence.

The caveats of multiple-intervention analysis are that it is more susceptible to the influence of heterogeneity compared to pair-wise analysis (51,52). There are several different models that can be run during a network meta-analysis to gauge the effect of heterogeneity. These statistical models include consistency, inconsistency and node-splitting models. If significant heterogeneity is detected, the inconsistency and node-splitting model results should be presented, and interpretations should be made with caution (53).

### Time-to-event data synthesis

In systematic reviews and meta-analyses, time-to-event outcomes such as survival are ideally pooled using meta-analysis of hazard ratios (HR). However, in many studies, the HR is not reported and individual patient data (IPD) is not available. Some studies have simply estimated actuarial survival outcomes visually from Kaplan-Meier plots and reported these in systematic reviews and meta-analysis. However, this approach does not take into account the censoring and loss of patients to follow-up that occurs, and does not allow estimation of a HR. To address this, several statistical approaches have been proposed in the literature that allows estimation of HR based on other summary statistics published (54-57). The estimated HR can then be used for meta-analysis to allow data synthesis from all available evidence in the literature.

One common HR estimation technique was proposed and validated by Tierney et al. in 2007 (58). In this technique, available data from Kaplan Meier curves can be digitized using computer software such as Digitizelt, and the recorded number of patients at risk noted from the paper. Therefore the computerized actuarial survival rate and number of patients at risk for sequential follow-up periods can be accurately estimated. Tierney et al. also developed

---

Figure 3 Meta-regression analysis of midpoint of study period versus paravalvular leak rate in Perceval S sutureless aortic valves. Significant negative correlation was found ($r = -0.853, P = 0.031$), suggesting that there was a decrease in paravalvular leak rate with time. This is one potential source of heterogeneity in the meta-analysis of proportions in Figure 1, suggesting that paravalvular leak is a function of temporal learning curve.
an Excel spreadsheet (58), when these data are inputted and assuming constant censoring, an estimation of the original IPD for the particular study can be deduced. Performing a similar analysis amongst all included studies for a meta-analysis can allow data synthesis of reconstructed IPD, thus facilitating meta-analysis of time-to-event outcomes.

IPD can also be reconstructed using another recent approach proposed by Guyot and colleagues in 2012 (59). This group has produced an iterative algorithm that is able to solve Kaplan-Meier equations used to produce the graphs in the original publication. Similar to methods by Tierney et al., a computerizing software such as Digitizelt can be utilized to digitize Kaplan-Meier curve data. This can then be entered into the iterative algorithm to determine optimal solutions to the Kaplan-Meier equations. Again, this algorithm assumes constant censoring and can be performed in the R statistical software. The reconstructed patient survival software can be aggregated to form combined survival curves. This approach was recently used to pool long-term time-to-event survival data for open surgical repair for chronic type B aortic dissection, producing aggregate Kaplan-Meier curves as shown in Figure 4 (60). This approach has also been recently used other reviews in the cardiothoracic surgery literature (61), and its usage is expected to increase in the near future with the greater need for pooled time-to-event data.

**Publication bias**

Another inherent limitation of systematic reviews is publication bias (62,63). Often in the literature, studies that have produced negative results are more difficult to publish compared to studies which produce a positive result. As a result, there is often more “missing” negative result studies, which may skew the outcomes of meta-analysis can provide misleading results.

As such, it is important that a systematic review or meta-analysis evaluates the potential influence of publication bias. One commonly used approach for the assessment of publication bias is the use of funnel plots (64). This is a plot that graphs precision versus magnitude of effect treatment and has a shape of an inverted funnel. The horizontal axis represents the intervention effect, and the vertical axis represents the standard error. Ideally, in the case of minimal publication bias, the points of the funnel plot are symmetrically distributed around the mean effect size. Asymmetrical distribution indicates potential publication bias which may undermine the validity of conclusions. Begg and Egger’s tests can be used in conjunction to statistically determine whether the asymmetry is significant or not (65).

To assess the effect of the “missing studies” due to publication bias on the calculated effect size, trim-and-fill analysis can be performed (66). This is an extension of the funnel plot technique, where “missing studies” are identified and estimated based on the symmetry of the funnel plot. The deduced missing studies can be imputed into the funnel plot, and the whether the resultant change in effect size is significant or not can assist in evaluation of publication bias. Recent meta-analyses in the cardiothoracic surgical literature have used this methodology to assess the effect of publication bias and “missing studies”, and an example is presented in Figure 5 (60,67,68). The caveat of this technique is that it is purely based on notion that no publication bias equates to a perfectly symmetry funnel plot, which may or may not be the case. Furthermore, the source or mechanism of publication bias is not addressed, and thus funnel plots and trim-and-fill analysis should be interpreted with caution.

**Interpretation of results**

There are several factors that must be considered when discussing results from the systematic review and meta-analysis. Firstly, the reviewers should assess the clinical significance of the results. For example, if there was a calculated significant difference in operation duration between two interventions as five minutes, would this significantly affect the outcomes of the patient? Prior
studies have suggested that the mean clinically important difference (MCID) of a therapy or intervention should be involved in the planning of clinical trials and interpretation of results (69). Secondly, the reviewer should discuss and explain potential sources of heterogeneity in the discussion. This may involve performing subgroup analysis and meta-regression analysis to determine which factors affect the outcome of interest. Thirdly, the review should discuss their strengths compared to prior reviews (if applicable) and limitations of the review, which may include but not limited to the inclusion of non-randomized studies (which may lead to risk of bias), small patient sample numbers, significant differences in baseline characteristics of comparative cohorts, short follow-up duration and heterogeneity in the surgical intervention techniques used amongst the included studies.

**GRADE assessment of outcomes**

The quality of scientific evidence and outcomes can further be assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (70,71). This is a transparent and structured approach for rating the evidence for a particular outcome. The GRADE methodology involves rating evidence for an outcome by upgrading or downgrading of evidence. Indications for upgrading evidence include having a large effect size and dose-response gradient. Indications for downgrading quality of evidence include serious risk of bias, serious inconsistency between studies, serious indirectness, serious imprecision and likely publication bias. The GRADE approach may allow the reviewer to have increased or decreased confidence in the effect sizes presented, that is, a higher confidence in the true association (71,72).

**Presentation of review results**

An overview of the systematic review and meta-analysis process is demonstrated in Figure 6. In order to assist surgeons in their rigorous conduction of systematic reviews...
and meta-analyses, we recommend the following structure for presentation of results (Table 2).

**Conclusions**

Systematic reviews and meta-analyses are increasingly important in the surgical realm for data synthesis and quality appraisal of available evidence. However, surgeons should be wary of the quality of systematic reviews, which may seriously undermine the validity of presented results and conclusions. In order to maintain high quality reviews with credible results, the review process should be standardized and strictly adhered. Here we have presented an overview of such a process to ensure optimal systematic review and meta-analysis outcomes and presentation.

**Acknowledgements**

*Disclosure:* The authors declare no conflict of interest.

**References**

11. Oxman AD, Guyatt GH. A consumer’s guide to subgroup...