Critical appraisal of systematic reviews

Abalos E, Carroli G, Mackey ME, Bergel E

Centro Rosarino de Estudios Perinatales, Rosario, Argentina

INTRODUCTION

In spite of the increasingly efficient ways to find, classify and store the information, as well as the availability of strategies aimed to extract and separate useful information from that which is neither valid nor applicable, clinicians are faced with the problem of the exponential expansion in the literature. This problem makes difficult, and sometimes impossible, to review all the articles addressing a clinical question, especially when dozens of articles reach conclusions that are relevant to the same question. Therefore, health providers, researchers, and policy makers need an efficient method to summarise the existing information in order to make evidence based decisions.

Traditional narrative reviews do not usually follow standardised and objective models. The decision to include or exclude articles depends to a large extent on the author's view, which in this type of reviews often does not take in account the methodological quality of the studies. Moreover, if there is no systematic search strategy of the literature, it is likely that the review does not include some important studies that could significantly alter its conclusions.
Systematic review: is a systematic search and critical evaluation of all primary studies answering the same question.
Meta-analysis: is the quantitative synthesis of all primary studies answering the same question using the same design.
Both systematic reviews and meta-analysis must be conducted following strict methodological rules, which must be described in detail to make them reproducible.

A systematic review aims to minimise the elements of arbitrariness in traditional narrative reviews, describing the review process in detail so, in principle, another person with access to the same resources could perform it arriving, in general, at the same results. This means that the objectives must be clearly established, the strategy for literature search documented and comprehensible, the evidence obtained subjected to a quality evaluation in an explicit way, and that the way in which evidence from individual studies is combined, clearly defined.

CRITICAL APPRAISAL

A systematic review is an exhaustive review of the literature addressing a clearly defined question, which uses a systematic and explicit methodology to identify, select and critically evaluate all the relevant studies, and collect and analyse the data emerging from the studies included in it. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the studies included in the review. These methods must be established beforehand and documented in the "Materials and methods" section. The meta-analysis involves the use of statistical techniques within a systematic review to combine the results of the included studies.

A systematic review establishes whether the scientific findings of research studies are consistent and whether these findings can be generalised to different populations, limiting the various possible forms of bias and increasing the reliability and precision of the estimates.

The purpose of the critical appraisal of a systematic review is to determine its validity, to interpret their results and to evaluate its applicability in clinical practice, in public health and/or in conducting future studies.

GUIDELINES FOR CRITICAL APPRAISAL

There are several questions guiding the reader through the process of critical appraisal of a systematic review. These could be grouped as validity of the review process, results and applicability of the results.

I. EVALUATION OF THE VALIDITY OF THE REVIEW

The first step for the critical appraisal of a review is to establish its methodological quality to determine the validity of the results. If the review has not been conducted with methodological rigour, it is unlikely that the results will reflect the truth and therefore they should not been taken into account, or they should be considered, but bearing in mind that the intrinsic errors may distort the results

1. Is the clinical question clearly focused with regard to:

- the population?
- the intervention?
- the outcome measures?

2. Are the criteria for the selection of the studies to be included in the review in accordance with:

- the specifications of the foregoing question in regard to populations, interventions and results?
- the type of research design that will be chosen?

3. Is the literature search method clearly specified?

- Is there a high probability that some relevant studies may have been omitted?

4. Have the identified studies been evaluated for methodological quality?

5. Was the methodological quality evaluation carried out by more than one person independently, and the degree of agreement between them established?

II. INTERPRETATION OF THE RESULTS OF THE REVIEW

Once you have evaluated the methodological quality of the review and reached the conclusion that reading the results is worthwhile, you must then interpret these results correctly. We shall discuss the appropriate techniques for quantitative interpretation of the results.

1. Were the results consistent from one study to another?

2. What were the overall results of the review?

3. How precise were the results?

III. APPLICABILITY OF THE RESULTS OF THE REVIEW IN CLINICAL PRACTICE

After having interpreted the results of the review we will need to evaluate whether they can be applied to our patients, and whether the benefits outweigh the potential harm.

1. Are my patients similar to the patients included in the original studies?

2. Is the intervention feasible in my setting?

3. Have all the clinically relevant results been taken into consideration?

4. Do the benefits outweigh the potential harm?

We will examine each of the above addressed questions in the context of the following clinical situation.

CLINICAL SITUATION

The practice of routine episiotomy for all vaginal deliveries is being discussed at a seminar in a hospital maternity unit. In this discussion it emerges that clinicians have different views on the policy that should be adopted. As personal views without any scientific support were raised, it was decided to make a search for systematic reviews on the topic of "episiotomy policies in vaginal births".

A systematic review titled Episiotomy policies in vaginal births" was found in The Cochrane Library, January 2001 issue. We will now work through this review which is also available in this issue of The WHO Reproductive Health Library.

I. EVALUATION OF THE VALIDITY OF THE REVIEW

1. Is the question the review is trying to answer clear and focused?

A systematic review must clearly establish the clinical question to which it relates, so that it can be determined whether it correctly states the purpose of the search. The same topic may give rise to different questions: preterm delivery may be evaluated in terms of prevention, aetiology, risk factors or management, and each of these

aspects could give rise to different systematic reviews. Only if the review establishes the objective pursued precisely it would be possible to determine whether the conclusions are relevant to the care of the patient in question.

The question should establish:

- Population: An intervention may be evaluated in different types of patients, selected by their age, stage of the disease, presence or not of risk factors, etc. The effects of the intervention may be different in these different groups. For example, the effect of an intervention in hospitalised patients may be different from the effect that would be shown in community-based studies.

The population included in the systematic review of episiotomy policies in vaginal births is described as pregnant women who had a vaginal delivery.

- Intervention: The type of intervention may be a pharmacological treatment or a surgical procedure, a diagnostic test or an exposure to a specific agent. Whatever the intervention, it must be stated clearly in the question. In practice, and in order to simplify interpretation of the text, we will limit ourselves to reviews that deal with evaluation of treatment effectiveness.

In our example, the main intervention compared is restrictive versus routine use of episiotomy, in general, also comparing restrictive versus routine use of mediolateral episiotomy, median episiotomy, and median versus mediolateral episiotomy.

- Outcomes: If an intervention is being evaluated, it should be specified what outcomes will be measured in order to establish its effect. Usually a primary objective is established fixing an outcome variable for a specific result, and adding other secondary outcomes. The variables for the outcomes to be measured should be those which are of clinical importance and crucial at the moment of deciding whether or not to give the intervention. The maternal and neonatal outcomes evaluated in the systematic review in our example are specified under the heading "outcomes", within the criteria for consideration of studies for review (see below).

2. Are the criteria for the selection of the studies clearly identified?

Selection criteria must be clearly established, as these criteria will determine which studies would be included in the systematic review and which ones would be excluded, and this will directly affect the results. These criteria should establish:

- Population: The specific characteristics of the patients in whom the intervention will be evaluated.

- Intervention: When the intervention concerns a form of treatment, the form of administration, dosing and duration of therapy should be specified so the reader can reliably determine the treatment that is being evaluated.

- Outcomes: The outcome variables by which the effect of the intervention will be measured should be specified.

- Methodological design: The type of epidemiological studies to be included should be specified. Randomized controlled trials provide the most reliable results with regard to the effectiveness of interventions. Therefore, to address such questions systematic reviews of randomized controlled trials should be conducted.

The criteria for inclusion in the episiotomy systematic review were as follows:

Population: Pregnant women who had a vaginal delivery.

Intervention: Primary: Restrictive versus routine use of episiotomy.

Secondary: Restrictive versus routine use of mediolateral episiotomy
Restrictive versus routine use of median episiotomy
Use of median episiotomy versus mediolateral episiotomy

Outcomes: Maternal: number of episiotomies, assisted delivery rate, severe vaginal/perineal trauma, severe perineal trauma, any posterior perineal trauma, any anterior trauma, need for suturing, estimated blood loss at delivery, perineal pain, use of analgesia, dyspareunia, perineal haematoma, healing complications, perineal wound dehiscence, perineal infection and urinary incontinence, and Neonatal: Apgar score below 7 at 1 minute, and admission to neonatal intensive care unit.

Type of studies: randomized controlled trials.

3. Is the literature search method specified?

"Is there a high probability that some relevant studies may have been omitted?" In a systematic review the literature search method used is of vital importance to ensure that the review is complete and updated. The search method must be explicitly stated so the reader can assess whether it has been carried out systematically and exhaustively, minimising the likelihood that relevant research has been omitted. The reviewer must ensure that the major part of the information of good quality that is available is included, whether or not it has been published or unpublished, and indexed in databases or not indexed, and that the search strategy is reproducible and documented in a clear and comprehensible manner.

Currently, a systematic search should include:

- An electronic archive of publications in general medicine, e.g. MEDLINE, EMBASE, LILACS, Cochrane Controlled Clinical Trials Register, etc., specifying the key words used and how they were used.

- Special databases focusing on the issue being addressed, eg. Popline for demographic studies.

- Review of the cited papers in the retrieved articles to look for further eligible articles and of the references of these articles in turn until this strategy is exhausted.

- Handsearch of publications that are specific to the question and perhaps not indexed in electronic databases.

- Personal communications with researchers or experts on the subject to identify unpublished articles, or to obtain data not included in the original publications.

- Informal discovery in discussions, conferences, congresses and correspondence.

The more these strategies are used, the more likely it is that relevant studies will not be omitted. These strategies ensure that there is little probability that relevant studies have been omitted. In our example, the authors used the search strategy developed by the Cochrane Collaboration Group, which includes all the items listed above. (This search strategy can be found in the Cochrane review).

4. Have the identified studies been evaluated for quality?

The review should proceed from the initial selection of the different studies to an evaluation of their quality in terms of their design, implementation and analysis to determine until what extent the results are reliable. In some cases the researcher may decide, in accordance with this quality evaluation, to exclude some of the studies because of important methodological shortcomings, and if so, he/she must explain the reasons for that.

There are many different quality scores to categorise articles in terms of their validity, but all of them are from data reports and have not been correctly validated, so that they should be viewed with caution. However, they may be useful in performing sensitivity analysis, which involves the analysis of the results excluding those studies of poor methodological quality to determine how they influence the overall results.

In our example, the authors list the quality evaluations to which the original articles were subjected:

- the control for selection bias at entry (the quality of random allocation assessing the generation and concealment methods applied),

- the control of selection bias after entry (the extent to which the primary analysis included every person entered into the randomised cohorts),

- intention-to-treat analysis: whether patients were analysed in the group to which they were assigned independently of the treatment received, and

- the control of bias in assessing outcomes (the extent to which those assessing the outcomes were kept unaware of the group assignment of the individuals examined).

The authors also specified which were the studies excluded and the reasons of their exclusions.

5. Was quality evaluation carried out by more than one person independently, establishing the degree of agreement between them?

Quality evaluation of studies should be carried out by more than one person using pre-established criteria. This is to minimise errors and confront differences in criteria for classification. This evaluation should be carried out independently, without knowing the names of the authors and the journals, the country of origin and the results, as this information could theoretically, influence the evaluation. The degree of agreement between the evaluators and the reasons for discrepancies should also be reported.

In the systematic review on episiotomy it is stated that two reviewers evaluated the quality of the studies independently, but the degree of agreement between them is not reported.

By answering the questions about methodological validity it is possible to determine the degree of reliability of the results of the systematic review and hence to decide whether it is worth reading it or not.

As the reader will have noted, the systematic review of policies on episiotomy in vaginal deliveries meets almost all the criteria for internal validity, so that it can be safely assumed that the results emerging from the review will reflect the true effect of the intervention, and it is unlikely that they could have been influenced by other factors.

II. INTERPRETATION OF THE RESULTS OF THE REVIEW

1. Were the results consistent from one study to another?

It is not uncommon that individual studies from a systematic review show different and even contradictory results. It is also relatively common to find that, although the results of individual studies agree on the efficacy of a therapy, they do not agree on the magnitude of its effect. Good reviews identify these differences and try to explain them. These differences in the results of the studies may be due to:

- Different populations: the populations included in the studies could differ with regard to certain characteristics which influence the outcome, for example, different stage or severity of the disease, different population characteristics such as age, sex, etc.

- Differences in the treatments administered: differences in dosing, route of administration or periods of treatment could alter the results between studies.

- Different ways of measuring the outcomes: the way in which the outcomes are measured may be different with regard to the technique, the frequency or the criteria used, which could jeopardise the comparability of results between studies.

- Different qualities of the studies: there is no doubt that the scientific methodological quality in which a study is conducted could modify its results and the differences between studies could be attributed to the differing qualities of the methods.

- The effect of chance: using a statistical test called the homogeneity test, it is possible to evaluate the probability that the differences between the results are exclusively due to chance and not to the factors mentioned above. Nevertheless, a clinical view of the differences may be more informative than the result of a hypothesis test as the differences may have no statistical significance but actually be of clinical importance, and vice versa.

In our example the authors of the systematic review carried out the necessary homogeneity tests, which were not statistically significant for most of the variables, and no clinically important inconsistencies were observed. They also carried out stratified analyses by type of episiotomy: mediolateral and median.

2. What were the results of the review?

In the meta-analyses a common measurement is calculated: the relative risk, which emerges from the weighted average of the relative risks of the studies included. The term weighted average is used because a greater weight is attributed to studies in which the variance of effect is small. The method used for its calculation is the Mantel-Haenzel-Peto. The interpretation of this result is the same as in the original studies.

When studies with different methodological qualities are included a sensitivity analysis must be carried out, which means calculating the results of the systematic review after taking out the studies of poor quality and observing how this affects the results. If the new result does not differ significantly from the overall result, it may be established that the studies of poor quality, and hence those more likely to give biased estimates, are not decisive with regard to the direction or magnitude of the effect of the treatment, which will give greater reliability in the results of the systematic review.

In the episiotomy study the estimated typical relative risk for severe perineal trauma was 0.80, which would represent a relative decrease in this event of 20%.

3. How precise were the results?

It is difficult that the point estimate of the effect of the treatment coincide exactly with its true value, since it is only an estimate based on a sample of patients (in this case all of the patients included in the original studies). If we want to know the range of values within which we can affirm with some confidence (usually 95%) that the estimated effect will occur in the general population, we must make use of the confidence intervals. The narrower the range included in the interval, the more precise the estimation of the result will be, and it will be possible to get a more reliable idea of the true effect of the treatment. The greater the number of patients experiencing the event of interest, the greater will be the precision of the estimated result. The sample size may be considered to be sufficient when the clinical conclusion about the efficacy of the treatment is the same for the whole range of values included in the confidence interval.

For example:

The values that are most likely to reflect the truth are those close to the estimated score, in the case of severe perineal trauma in the episiotomy study: 0.80, becoming

less probable as the limits of the confidence interval are approached. The 95% confidence intervals for severe perineal trauma are: 0.55-1.16.

The results of each study are usually presented in graphic form, together with their respective confidence intervals. A black square and a horizontal line represent each study, which represent the point estimate and the confidence interval for this relative risk, respectively. The size of the black square represents the weight of the study in the meta-analysis. A solid vertical line corresponds to no effect of the treatment (relative risk=1.0). If the line representing the confidence interval includes the value 1, the difference in the effect between the control group and the experimental group is not significant within established limits (95%). The diamond represents the combination (using the weighted average) of the results of the primary studies. The horizontal edges of the diamond represent the limits of the confidence interval. The authors of the review in our example present the results as relative risks with their respective confidence intervals.

III. APPLICABILITY OF THE RESULTS OF THE REVIEW IN CLINICAL PRACTICE

1. Are my patients similar to the patients included in the original studies?

Before deciding whether the effectiveness of the treatment shown by the review will apply to our patients, we must determine whether our patients are similar to those included in the review. To do so, we should ask whether there is any reason that the treatment might have a different effect in our patients on account of physiological or clinical characteristics, co-morbid factors or specific contraindications.

It cannot always be strictly deduced from a study that the evidence provided by it can or should be applied to any particular patient. The results of different studies are usually presented as average effects, and patients may differ from this average, and, thereby influencing the effectiveness of the treatment (reduction of relative risk) or its impact (reduction of absolute risk). The patients participating in the research studies may not be the same as the type of patients in whom the treatment could be potentially useful. Nevertheless it is probably more appropriate to assume that the results of a systematic review can be generalised to all patients unless there is strong theoretical or empirical evidence suggesting that a particular group may respond differently.

There may be heterogeneity in the effects between different patients due to biological, social or other differences which influence the effect of the intervention or the risk of an adverse result. Doctors and patients should consider the following points before applying evidence from a review to a particular case:

- Whether the results may be different for this case because the patient's physiological or clinical characteristics make her different from the patients included in the review.

- What is the absolute risk of an adverse effect for this patient if the treatment is not applied.

- Whether this patient has significant co-morbid factors or contraindications that may reduce the benefit reported in the review.

- Whether there are social or cultural factors that may affect the feasibility or acceptance of the treatment.

- What is the decision of the patient or her relatives.

Clinicians may continue casting doubts about the applicability of the results because of small differences in the characteristics of their patients, or because the review shows results taken from a combination of studies that use different derivatives of a generic drug and they would like to find out whether one has greater effectiveness than the others. These questions raise the issue of subgroup analysis. However, it is necessary to bear in mind certain criteria that must be fulfilled in order to have confidence in the results of subgroup analysis. These are:

- Statistically significant difference in the overall effect of the treatment makes subgroup analysis easier to justify. It is more appropriate to conduct a subgroup analysis when the overall results are statistically significant rather than the other way around.

- Previously established hypothesis, being one of the few that were tested.

- Consistency throughout the studies.

- Biological plausibility.

Due to the characteristics of the patients included in the six studies covered by the systematic review of episiotomy policies (single pregnancies, term, cephalic presentation, no contraindications for vaginal delivery) we infer that the results conclusions are applicable to our population.

2. Is the intervention feasible in my environment?

It is impossible to actively promote implementation of the results of all systematic reviews because of the limited capacity of health systems to absorb the new evidence and implement the necessary measures to do away with the obstacles that hamper translation of the results of theory into practice. So when the clinician is looking at the results of a study he/she must consider the feasibility of applying them in the environment in which he/she is working, taking into account the technical factors and infrastructure, as well as such factors as the training of staff, if required, changes in established practice in the service, and also the acceptance by patients or their relatives that this new intervention should be applied to them.

In the example with which we are concerned, where the usual practice in many services is the routine use of episiotomy in all vaginal deliveries, the implementation of a policy of restrictive use of this practice, in the light of the results of the available evidence, would not make it more difficult to carry out the usual activities of the service.

3. Have all the clinically important results been taken into consideration?

Although the results of reviews are generally focused on the final primary points raised in the hypothesis, the secondary points that are clinically and biologically

important must also be considered, as well as reports of side-effects or adverse events.

If the review does not include the outcomes that you think are important, you should refer back to the original studies and check whether this event has been taken into account, and what is the effect of the intervention for this result. If, for example, we want to evaluate the effectiveness of aspirin to prevent myocardial infarction in patients with chronic stable angina, it is necessary to look at what happens with mortality and severe haemorrhage, as both these results are fundamental before deciding whether or not to administer this treatment.

The systematic review on episiotomy policies, in addition to the final main outcome concerning severe perineal trauma (grade III and IV), looks at outcomes such as types of vaginal/perineal trauma (anterior, posterior), the need for suturing, estimated blood loss, perineal pain, need for analgesia, dyspareunia, presence of haematomas, infections, urinary incontinence, and variables affecting the newborn, such as Apgar score and admission to a neonatal intensive care unit. Many of these variables are present in general and in subgroup analysis taking in account parity and type of episiotomy (median or mediolateral).

4. Do the benefits outweigh the potential harm and costs?

Finally, when a clinical decision is taken, the expected benefits should be evaluated in relation to potential harm. In a treatment having known adverse side-effects it must be assessed whether it is justified to subject the patient to these effects with the aim of achieving the expected benefit. When the evidence on itself is examined, those responsible for evaluating it and drawing conclusions about it may differ as to the criteria for establishing priorities for implementation. Health policy-makers, for example, may be looking for social benefits in health and efficacy, while clinicians may regard the welfare of their individual patients as their most important goal. This is the point at which the concept of number-needed-to-treat (NNT) becomes extremely important. The NNT is defined as the number of patients needed to treat to avoid an additional adverse event or prevent an additional complication of the disease, and it is calculated by taking the inverse of the absolute risk reduction (1/ARR).

In the episiotomy example, the authors found that 45 of 1,843 patients suffered third and fourth degree tear (severe perineal trauma) in the group in which episiotomy was selectively used (intervention group), as compared with 56 of the 1,807 patients in whom episiotomy was routinely used (control group). It can thus be seen that the incidence of the problem was 2.4% (45/1,843) in the intervention group and 3.1% (56/1,807) in the control group. The risk of severe perineal tears was reduced by 0.7% (ARR= [56/1807 - 45/1843] x 100), needing to treat 143 patients to prevent one of these events (1/ARR, or 1/0.007).

NNT takes into account not only the reduction of the risk evidenced by the proposed treatment, but also relates it to the incidence of the problem. Thus, even a very high reduction of risk may not have a very great impact on the population if the incidence of the problem is very low, as we would have to treat many patients to prevent just one event, and sometimes this is not justified specially when treatments are not always harmless, cheap or easy to administer. On the other hand, another treatment which demonstrates a moderate or even small reduction in the risk of suffering an event but is applied to populations where the incidence of the problem is high may

show significant clinical results because it will be necessary to treat only a few patients to avoid an unfavourable result.

This document should be cited as: Abalos E, Carroli G, Mackey ME, Bergel E. Critical appraisal of systematic reviews: The WHO Reproductive Health Library, No 4, Geneva, The World Health Organization, 2001 (WHO/RHR/01.6).