3.2: Systematic reviews

Reviewing the literature can be a daunting task. The volume of information available through published papers, or the Internet, is vast and constantly expanding. Given the volume of literature available, an ‘ad hoc’ review of the literature is subject to substantial biases if only some studies are included, since the studies that are found this way may well not be representative of all the relevant studies. The best way to ensure an objective and unbiased review of the literature is to conduct a review that follows strict guidelines to minimize bias in selecting and interpreting reported studies.

The basic steps in a systematic review are shown in Box 3.1.

In this chapter, we provide a brief overview of each of these steps. Further details are given in published guidelines, such as the Cochrane handbook for systematic reviews of interventions (Higgins and Green, 2008) and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (<http://www.prisma-statement.org>) (Liberati et al., 2009), and books on systematic reviews in health research (Egger et al., 2001, Glasziou, 2001, Khan, 2003).

Box 3.1 The five basic steps in a systematic review

1. Defining the question.
2. Identifying relevant studies in a predefined, systematic way.
3. Assessing the quality of each relevant study.
4. Summarizing the evidence.
5. Interpreting the findings.
2.1 Defining the question

The first step in a systematic review is to define the research question. A structured approach for framing the question is useful—the PICOS approach (Population; Interventions (or Exposure); Comparison; Outcomes; Study design) (Higgins and Green, 2008) is used by both Cochrane and PRISMA.

For example, a systematic review summarized the evidence of the effectiveness of behavioural interventions to prevent HIV infection among young people in sub-Saharan Africa (Napierala Mavedzenge et al., 2011). The review question was structured, using the PICOS approach, as follows:

- **Population**: Among young people aged 10–24 years in sub-Saharan Africa . . .
- **Intervention/exposure/comparison**: . . . does exposure to an intervention focusing on reducing HIV risk behaviours, relative to no or minimal intervention, . . .
- **Outcomes**: . . . reduce the risk of HIV, STIs, or pregnancy . . .
- **Study design**: . . . when evaluated through experimental or quasi-experimental study designs?

A second example, used in this chapter, is a systematic review of the evidence that the use of chewing substances (such as smokeless tobacco or betel nuts) is associated with cardiovascular disease (CVD) in Asia (Zhang et al., 2010). In this case, the question was structured as follows:

- **Population**: Among people in Asian countries . . .
- **Intervention/exposure/comparison**: . . . does exposure to chewing substances, relative to not chewing them, . . .
- **Outcomes**: . . . increase the risk of CVD . . .
- **Study design**: . . . when evaluated through observational epidemiological studies?

Previous systematic reviews had examined this question in the United States of America (USA) and Sweden, but there was no synthesis of the evidence from Asia. If strong evidence for an association was found, this could lead to the development and evaluation of an intervention directed at reducing betel chewing in these populations.

Once the research question is identified, a detailed protocol should be prepared for the review. This will include definition of the search strategy and the planned analyses. There are plans to develop an international register of systematic reviews, led by the Centre for Reviews and Dissemination (<http://www.york.ac.uk/inst/crd/index.htm>), which will enable researchers to register their review protocol. This will extend the register developed by the Cochrane Collaboration (<http://www.cochrane.org>), which was established in 1993 to promote systematic reviews of health care interventions. Researchers undertaking reviews under the Cochrane Collaboration are required to register the protocol for their review in advance, and the review is peer-reviewed before publication. However, many systematic reviews are undertaken outside of the Collaboration and may not currently be registered.

2.2 Identifying relevant literature

to be searched (Higgins and Green, 2008). However, these databases have a North American/European bias, and, for studies in LMICs, it is worth also searching other relevant databases such as LILACS (Latin American Caribbean Health Sciences Literature), African Healthline, GlobalHealth, and Popline. In addition, there are many subject-specific databases, such as PsychInfo (for psychology and related behavioural and social sciences), as well as Internet search engines such as Google Scholar. It may also be useful to search conference databases and trial registries to identify additional papers.

Strategies can be used to identify both free-text words in the database and controlled terms (called MeSH in MEDLINE, i.e. medical subject headings) that are used as keywords. Search strategies need to include the key terms in the review question and use the Boolean operators (such as 'AND', 'OR', 'NOT') to produce a search that is both sensitive and specific to the research question. The search strategy used for the example of chewing substances and CVD in Asia is given in Box 3.2.

Often the reviewers will already know about some key published studies. It is useful to check that all of these have been identified by the electronic database search. If not, a careful review of the search strategy may establish the reason for this, and the search can be amended accordingly.

Table 3.1 Inclusion criteria: example for the systematic review of behavioural interventions to prevent HIV infection among young people in sub-Saharan Africa

<table>
<thead>
<tr>
<th>PICOS component (see text)</th>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Young people aged 10–24 years. In studies with a wider age range, there must be an analysis of the impact of the intervention in young people (10–24 years) or, at least, in part of that age range. In sub-Saharan Africa. Based in a school, and/or health facility, and/or geographically defined community.</td>
<td>Study population not representative of a general population of young people (for example, young sex workers). Fewer than 100 people in the study.</td>
</tr>
</tbody>
</table>
| Intervention/exposure     | Behavioural intervention focused on one or more of the following:  
(i) improving sexual and reproductive health skills and behaviour  
(ii) reducing the risk of sexually transmitted diseases (STDs)  
(iii) reducing unintended pregnancies  
(iv) increasing utilization of health services for treatment of STIs and/or behaviours related to more appropriate service utilization. | No suitable comparison group (for example, non-randomized study with post-intervention data) |
| Comparison                | No or minimal behavioural intervention. | |

https://med.libretexts.org/Bookshelves/Nursing/Book%3A_Field_Trials_of_Health_Interventions_-_A_Toolbox_(Smith_Morrow…
Updated: Fri, 23 Sep 2022 17:43:26 GMT
Powered by
<table>
<thead>
<tr>
<th>PICOS component (see text)</th>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>only). No adjustment for differences between groups that might bias the findings.</td>
</tr>
</tbody>
</table>

**Outcome**

At least one of the following measured:

- (i) prevalence or incidence of HIV infection
- (ii) prevalence or incidence of another STI
- (iii) prevalence or incidence of pregnancy (measured by laboratory test or clinically observed)
- (iv) reported sexual and reproductive health behaviour (including treatment-seeking behaviour).

Measured less than 3 months after the intervention starts.

**Study design**

Published in 2005–2008 (because an earlier systematic review had covered the period up to the end of 2004). Randomized and non-randomized epidemiological studies which included a contemporaneous comparison group or a before–after/time series analysis in the intervention group only.

---

Box 3.2 Example of a search strategy for evidence of an association between chewing substances and CVD, ischaemic heart disease, or cerebrovascular disease in Asia

We searched PubMed (up to July 2010), using the terms: (‘cardiovascular diseases’ [MeSH] OR ‘cardiovascular’ [All Fields] AND ‘diseases’ [All Fields]) OR ‘cardiovascular diseases’ [All Fields] OR ‘cerebrovascular disorders’ [MeSH] OR (‘cerebrovascular’ [All Fields] AND ‘disorders’ [All Fields]) OR ‘cerebrovascular disorders’ [All Fields] OR ‘stroke’ [MeSH] OR ‘stroke’ [All Fields] OR ‘mortality’ OR death*) AND (‘betal quid’ OR ‘betal-quid’ OR ‘betal nut’ OR ‘betal nuts’ OR ‘areca nut’ OR ‘areca nuts’ OR ‘paan’ OR ‘pan’ OR ‘snuff’ OR ‘snus’ OR ‘gul’ OR ‘gutka’ OR ‘khaini’ OR ‘loose leaf’ OR ‘maras’ OR ‘mawa’ OR ‘mishri’ OR ‘naswar’ OR ‘Areca catechu’ OR ‘tooth powder’ OR ‘shammah’ OR ‘tobacco chewing gum’ OR ‘zarda’ OR ‘tobacco, smokeless’ [MeSH] OR ‘smokeless tobacco’ OR ‘chewing tobacco’ OR ‘non-smoking tobacco’) AND (‘cohort studies’ [MeSH] OR ‘cross-sectional studies’ [MeSH] OR ‘case control studies’ [MeSH] OR (‘cohort’ [TI] AND stud* [TI]) OR (case* [TI] AND control* [TI]) OR ‘prospective’ OR ‘retrospective’ OR ‘cross-sectional’ OR ‘cross sectional’), which yielded 1006 potentially relevant references. We adapted the searching strategy for a second search in ISI Web of Science (updated 19 July 2010) and found another 739 references. We identified all observational studies, including cohorts, case-control studies, and cross-sectional studies, provided that they explored the association between ever using chewing substances and the occurrence (incidence or mortality) of CVD and reported the strength of the associations with a quantitative risk estimate. There was no limitation on the language, study year, or publication status.


https://med.libretexts.org/Bookshelves/Nursing/Book%3A_Field_Trials_of_Health_Interventions_-_%26_A_Toolbox_(Smith_Morrow…
2.2.2 Reviewing abstracts

The search strategy commonly identifies several thousands of potentially relevant papers. The next step is for two reviewers to independently read through the abstract of each paper and define it as being potentially relevant or not. At this stage, it is recommended to err on the side of caution, i.e. include as ‘potentially relevant’ if the relevance is unclear from the abstract. The two reviewers should then compare their results and reconcile any differences by discussion, further reference to the abstracts, or a third reviewer independently reading the abstract.

2.2.3 Reviewing full articles

Full copies of all papers, the abstracts of which were considered to be potentially relevant, should be obtained (electronically, from libraries, or by emailing the author). They should be reviewed by the two reviewers who independently assess whether or not each paper meets each of the inclusion/exclusion criteria. Discrepancies should be resolved as for the abstracts.

2.2.4 Hand searching

The next step in the search strategy is usually to review the reference lists of all the eligible studies identified from the electronic database search, to identify any studies that were missed by that search but have been referenced in the eligible papers.

Previous review papers should also be read to check that no known papers have been omitted. Finally, it is legitimate, though sometimes time-consuming, to include unpublished studies which can be identified through colleagues or contact with the investigators of unpublished studies, for example, identified through Internet searches or trial registers. It is also important to identify ongoing studies, where possible, as these may be included in updates of the review.

2.2.5 Flow chart of search strategy

The template for a flow chart summarizing the search results is given in Figure 3.1. In the example of behavioural interventions among young people in sub-Saharan Africa, a total of 1173 papers were identified from the electronic databases, of which 137 were deemed potentially relevant after review of their titles and abstracts, and full-text articles were obtained. After excluding those not meeting the inclusion criteria, the final review included 40 papers, representing 23 studies (as sometimes the results of one study were reported in more than one paper) (Napierala Mavedzenge et al., 2011). For the example of chewing substances and CVD in Asia, 1756 publications were identified from electronic databases, of which only six were eligible for inclusion in the analysis of CVD (Zhang et al., 2010).

2.3 Descriptive synthesis of studies

When the eligible papers have been identified, a data extraction form should be completed for each study, which contains fields enabling a detailed description of the study design and of the results. For example, descriptive elements...
would include the PICOS components, as discussed in Section 2.1. The results should focus on the pre-specified outcomes in the review protocol and would include outcome measures, definition of exposures/interventions, measures of effect, and 95% confidence intervals (CIs). The form should be pilot-tested on a few sample papers and revised, as appropriate. Two reviewers then read each paper in detail independently, summarize the paper on to the data extraction form, and appraise the risk of biases. A common shortcut, which is permissible, is that one reviewer completes the data extraction form and the other then checks and edits it, with the final version based on a discussion of any discrepancies.

The next step is to begin to summarize the evidence from the eligible studies as a whole. All reviews should include a descriptive table of the included studies, which summarize the study population, intervention, comparison, outcome, and study design. One of the 23 studies that were identified in the review of behavioural interventions among young people is summarized in Table 3.2.

In the table that summarizes the results of each study, all the primary and secondary outcome measures should be included. For a binary outcome, this would include the proportion with the outcome among the exposed and unexposed groups, the appropriate measure of effect (e.g. risk ratio (RR), rate ratio (RR), or odds ratio (OR)), and 95% CI. For continuous outcomes, the mean, standard deviation in the exposed and unexposed, plus the effect measure (e.g. standardized mean difference) should be given.

From Moher et al., Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, *PLoS Medicine*, Volume 6, Issue 7, e1000097, Copyright © Moher et al. 2009. This figure is reproduced from an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Table 3.2 Description of one of the studies included in the systematic review of youth interventions against HIV infection in sub-Saharan Africa
<table>
<thead>
<tr>
<th>Study, location, and programme</th>
<th>Type of intervention and setting</th>
<th>Target population, primary objectives, comparison, and study outcomes</th>
<th>Intervention description</th>
<th>Study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Republic of Tanzania, MEMA kwa Vijana</td>
<td><strong>Schools:</strong> Teacher-led. Curriculum-based sexual and reproductive health education. <strong>Health facility:</strong> Interventions to facilitate youth friendliness of service providers, linked to interventions in the community and in other sectors (schools), to promote acceptance and utilization.</td>
<td><strong>Target population:</strong> Persons aged 12–19 years in rural areas. <strong>Primary objectives:</strong> Delayed sexual initiation, increased condom use, decreased number of sexual partners, and increased use of health services, especially for sexual and reproductive health services. <strong>Comparison arm:</strong> Current (very limited) sexual and reproductive health education in schools, and no additional interventions within health facilities or in the wider community. <strong>Study outcomes:</strong> <strong>Primary:</strong> HIV incidence; HSV2 prevalence. <strong>Secondary:</strong> pregnancy (by test and self-reported); prevalence of other STIs (by test and self-reported); knowledge and attitudes related to sexual and reproductive health issues; self-reported sexual risk behaviours, including sexual debut during trial follow-up, use of condoms, number of sexual partners, use of health services if reported a potential STI.</td>
<td>In-school teacher-led and peer-assisted programme. Covered refusal, self-efficacy, self-esteem, STI/ HIV, sexuality, contraception, social values, respect, gender. Used drama, stories, and games. Also included interventions to make government health services more youth-friendly, youth condom promotion and distribution, and limited community-wide interventions. Ten to 15 lessons per year over 3 years.</td>
<td>-</td>
</tr>
</tbody>
</table>

Adapted with permission from *Journal of Adolescent Health*, Volume 49, Issue 6, Napierala Mavedzenge et al., HIV prevention in young people in sub-Saharan Africa: a systematic review, pp. 568–86, Copyright © 2011 Society for Adolescent Health and Medicine. Published by Elsevier Inc. All rights reserved. [http://www.sciencedirect.com/science/journal/1054139X]. This table is not covered by the Creative Commons licence terms of this publication. For permission to reuse please contact the rights holder.

### 2.4 Assessing risk of bias in the studies

Once the description of each study is completed, an evaluation should be conducted of the extent of potential bias and error that may have arisen, either from the design or the analysis of each of the original studies. The main aim of this is to guide interpretation of the findings of the review. In some cases, it may be decided to exclude a study which is flawed to the extent that the results are considered likely not to be valid. Alternatively, a sensitivity analysis might be conducted to evaluate how the summary results differ if results from more flawed studies are included or excluded.

There are several methods for assessing the risk of bias, including checklists or ‘quality score’ scales. The recommendation of the Cochrane Collaboration and the PRISMA guidelines is to use a ‘domain-based evaluation’, in
which critical assessments are made for domains such as blinding of participants and generation of the random sequence (for randomized studies) (Higgins and Green, 2008). For observational studies, there are additional possible sources of bias. For example, in case-control studies, check should be made on the external validity of case selection, the choice of control group, and adjustment for confounding factors.

Table 3.3 summarizes some of the sources of potential bias in RCTs and observational studies.

The assessment of potential biases should be tailored to the research question. For each review, there should be consideration of whether one potential bias is more important to the interpretation of findings than others. For example, if an outcome is measured objectively (for example, mortality), then blinding of those evaluating the outcome is not going to be very important. In contrast, if loss to follow-up is high and associated with the outcome, then this could cause substantial bias.

A table summarizing the risk of bias in each study should be completed independently by two reviewers, and any differences reconciled by discussion or reference to a third reviewer. Summarizing the results can be done in different ways—some authors rank the studies in order of quality; others divide them into those with low, medium, or high risk of bias. These decisions should be taken independently of the results of the studies, if possible, before examining the results, and the reviewers need to decide which studies (if any) will be taken forward to a quantitative meta-analysis of findings.

Table 3.3 Methods for assessing risk of bias in RCTs and observational studies

<table>
<thead>
<tr>
<th>Source of bias</th>
<th>Definition</th>
<th>Assessment for RCTs</th>
<th>Assessment for observational studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Systematic differences between the comparison groups</td>
<td>Generation of random allocation concealment</td>
<td>Selection of exposed/ unexposed Selection of cases/ controls</td>
</tr>
<tr>
<td>Performance bias</td>
<td>Systematic differences in the care provided (apart from intervention)</td>
<td>Blinding of participant and provider Misclassification of exposure</td>
<td>Systematic differences in those exposed and unexposed Misclassification of exposure</td>
</tr>
<tr>
<td>Attrition bias</td>
<td>Systematic differences between the comparison groups in withdrawals from the study</td>
<td>Intention-to-treat analysis Outcome data not available for all participants</td>
<td>Differing follow-up rates between exposed and unexposed (or participation rates in cases and controls)</td>
</tr>
<tr>
<td>Detection bias</td>
<td>Systematic difference in outcome assessment</td>
<td>Blinding of those evaluating outcome</td>
<td></td>
</tr>
</tbody>
</table>

2.5 Quantitative synthesis of results
2.5.1 Forest plots

Following the descriptive analysis and assessment of risk of bias, it may or may not be appropriate to conduct a formal meta-analysis that quantifies the overall effect of the intervention. If, for example, the study populations, interventions, and reported outcomes differed substantially, the authors may decide to focus on describing the studies, their results, applicability, and limitations in a narrative review, rather than produce a quantitative summary. This was the case for the systematic review of interventions in young people in sub-Saharan Africa (Napierala Mavedzenge et al., 2011).

In other cases, it might be useful to summarize the data quantitatively. A first step for this is to produce a graph, called a forest plot, which displays the measure of effect (e.g. OR) for each study, together with a horizontal line denoting the CI. Before constructing such a graph, it is important to consider whether the results from the different studies are indeed measuring the same effect and are comparable to each other. For example, a smoking cessation intervention may have a different effect in pregnant women than among teenage girls. In such cases, it would be beneficial to present results stratified by subgroups, in whom effects might be expected to differ. As with all analyses, these subgroups should be defined in advance and included in the review protocol. For example, in the review of chewing substances in Asia, it was decided a priori to stratify by geographical region, to minimize confounding due to the presence or absence of tobacco in chewing substances, as this was thought to differ between regions.

In this example, the six eligible studies included five cohort studies and one case-control study. The forest plot is shown in Figure 3.2. The solid vertical line indicates a relative risk (RR) of one, representing no association between the exposure and outcome. In this example, all six studies had a RR greater than one, indicating an increased risk of CVD among individuals who used chewing substances, and the 95% CI did not include one for four of these studies, indicating strong evidence of an association. The forest plot also includes an overall (summary) estimate of the RR. This is a weighted average of the effects from each of the studies.

There are two main methods of obtaining the summary measure of an intervention effect. In a ‘fixed-effects’ model, it is assumed that the true effect of exposure (or the intervention) is the same in each study, any variation between studies being solely due to chance. In contrast, a ‘random-effects’ model may be used, in which the true effect of exposure for the individual studies are assumed to inherently vary (e.g. due to differences in the populations or residual confounding factors). In a random-effects model, the weights allow for this between-study variation, as well as the random variation.

In Figure 3.2, a random-effects model was used, and the weights for each study are given on the right-hand side of the forest plot. The overall (summary) estimate is RR = 1.26, with a 95% CI of 1.12–1.40. Note that this summary estimate is more precise (i.e. has a narrower CI) than any one of the individual studies. By undertaking a systematic review and meta-analysis, the reviewers can now report that there is strong evidence that, in these populations, exposure to chewing substances was associated with an increased risk of CVD of around 26%, compared with non-users.
2.5.2 Examining heterogeneity

The effect sizes of individual studies will inevitably be different from each other, but it is important to assess whether this difference is likely to be due to random variation (i.e. the true underlying effect will be the same) or to real differences in underlying effect sizes in the individual studies. It is therefore essential to examine the consistency of the effects and to quantify the heterogeneity (or difference) in effect sizes between studies. Several measures are available for this, one of which is the $I^2$ statistic (Higgins et al., 2003). This statistic is the percentage of total variation across studies that is due to heterogeneity, rather than chance. A value of $I^2$ of 0% indicates no observed heterogeneity, and larger values indicate increasing heterogeneity. The principal advantage of the $I^2$ statistic is that it does not depend on the number of studies included in the meta-analysis and so can be used even for meta-analyses containing relatively few studies, which typically have low power to detect heterogeneity using other measures.

In our example, the value of $I^2$ is 35.9%, with a p-value of 0.17, indicating little evidence of heterogeneity. The reviewers were therefore justified in presenting the summary estimate. If, in contrast, the $I^2$ statistic suggests evidence of heterogeneity, for example if $I^2$ was 70%, further exploration of the causes of heterogeneity would be needed, for example by undertaking (pre-specified) subgroup analyses. If there was no longer evidence of heterogeneity within subgroups, this would indicate that the stratifying characteristics were an important source of heterogeneity, and results should be presented within subgroups, rather than overall.