Procedures for Performing Systematic Reviews

Barbara Kitchenham

e-mail: barbara@cs.keele.ac.uk

Joint Technical Report

Software Engineering Group
Department of Computer Science
Keele University
Keele, Staffs
ST5 5BG, UK

Keele University Technical Report TR/SE-0401
ISSN:1353-7776

and

Empirical Software Engineering
National ICT Australia Ltd.
Bay 15 Locomotive Workshop
Australian Technology Park
Garden Street, Eversleigh
NSW 1430, Australia

NICTA Technical Report 0400011T.1

July, 2004

0. Document Control Section

0.1 Contents

0.2 Document Version Control

Document status	Version Number	Date	Changes from previous version
Draft	0.1	1 April 2004	None
Published	1.0	29 June 2004	Correction of typos Additional discussion of problems of assessing evidence Section 7 “Final Remarks” added.

0.3 Executive Summary

The objective of this report is to propose a guideline for systematic reviews appropriate for software engineering researchers, including PhD students. A systematic review is a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest. Systematic reviews aim to present a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

The guideline presented in this report was derived from three existing guidelines used by medical researchers. The guideline has been adapted to reflect the specific problems of software engineering research.

The guideline covers three phases of a systematic review: planning the review, conducting the review and reporting the review. It is at a relatively high level. It does not consider the impact of question type on the review procedures, nor does it specify in detail mechanisms needed to undertake meta-analysis.

1. Introduction

This document presents a general guideline for undertaking systematic reviews. The goal of this document is to introduce the concept of rigorous reviews of current empirical evidence to the software engineering community. It is aimed at software engineering researchers including PhD students. It does not cover details of meta-analysis (a statistical procedure for synthesising quantitative results from different studies), nor does it discuss the implications that different types of systematic review questions have on systematic review procedures.

The document is based on a review of three existing guidelines for systematic reviews:

1. The Cochrane Reviewer's Handbook [4].
2. Guidelines prepared by the Australian National Health and Medical Research Council [1] and [2].
3. CRD Guidelines for those carrying out or commissioning reviews [12].

In particular the structure of this document owes much to the CRD Guidelines.

All these guidelines are intended to aid medical researchers. This document attempts to adapt the medical guidelines to the needs of software engineering researchers. It discusses a number of issues where software engineering research differs from medical research. In particular, software engineering research has relatively little empirical research compared with the large quantities of research available on medical issues, and research methods used by software engineers are not as rigorous as those used by medical researchers.

The structure of the report is as follows:

1. Section 2 provides an introduction to systematic reviews as a significant research method.
2. Section 3 specifies the stages in a systematic review.
3. Section 4 discusses the planning stages of a systematic review
4. Section 5 discusses the stages involved in conducting a systematic review
5. Section 6 discusses reporting a systematic review.

2. Systematic Reviews

A systematic literature review is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest. Individual studies contributing to a systematic review are called *primary* studies; a systematic review is a form a *secondary* study.

2.1 Reasons for Performing Systematic Reviews

There are many reasons for undertaking a systematic review. The most common reasons are:

- To summarise the existing evidence concerning a treatment or technology e.g. to summarise the empirical evidence of the benefits and limitations of a specific agile method.

- To identify any gaps in current research in order to suggest areas for further investigation.
- To provide a framework/background in order to appropriately position new research activities.

However, systematic reviews can also be undertaken to examine the extent to which empirical evidence supports/contradicts theoretical hypotheses, or even to assist the generation of new hypotheses (see for example [10]).

2.2 The Importance of Systematic Reviews

Most research starts with a literature review of some sort. However, unless a literature review is thorough and fair, it is of little scientific value. This is the main rationale for undertaking systematic reviews. A systematic review synthesises existing work in manner that is fair and seen to be fair. For example, systematic reviews must be undertaken in accordance with a predefined search strategy. The search strategy must allow the completeness of the search to be assessed. In particular, researchers performing a systematic review must make every effort to identify and report research that does not support their preferred research hypothesis as well as identifying and reporting research that supports it.

2.3 Advantages and disadvantages

Systematic reviews require considerably more effort than traditional reviews. Their major advantage is that they provide information about the effects of some phenomenon across a wide range of settings and empirical methods. If studies give consistent results, systematic reviews provide evidence that the phenomenon is robust and transferable. If the studies give inconsistent results, sources of variation can be studied.

A second advantage, in the case of quantitative studies, is that it is possible to combine data using meta-analytic techniques. This increases the likelihood of detecting real effects that individual smaller studies are unable to detect. However, increased power can also be a disadvantage, since it is possible to detect small biases as well as true effects.

2.4 Feature of Systematic Reviews

Some of the features that differentiate a systematic review from a conventional literature review are:

- Systematic reviews start by defining a review protocol that specifies the research question being addressed and the methods that will be used to perform the review.
- Systematic reviews are based on a defined search strategy that aims to detect as much of the relevant literature as possible.
- Systematic reviews document their search strategy so that readers can access its rigour and completeness.
- Systematic reviews require explicit inclusion and exclusion criteria to assess each potential primary study.
- Systematic reviews specify the information to be obtained from each primary study including quality criteria by which to evaluate each primary study.
- A systematic review is a prerequisite for quantitative meta-analysis

3. The Review Process

A systematic review involves several discrete activities. Existing guidelines for systematic reviews have different suggestions about the number and order of activities (see Appendix 1). This documents summarises the stages in a systematic review into three main phases: Planning the Review, Conducting the Review, Reporting the Review.

The stages associated with *planning the review* are:

1. Identification of the need for a review
2. Development of a review protocol.

The stages associated with *conducting the review* are:

1. Identification of research
2. Selection of primary studies
3. Study quality assessment
4. Data extraction & monitoring
5. Data synthesis.

Reporting the review is a single stage phase.

Each phase is discussed in detail in the following sections. Other activities identified in the guidelines discussed in Appendix 1 are outside the scope of this document.

The stages listed above may appear to be sequential, but it is important to recognise that many of the stages involve iteration. In particular, many activities are initiated during the protocol development stage, and refined when the review proper takes place. For example:

- The selection of primary studies is governed by inclusion and exclusion criteria. These criteria are initially specified when the protocol is defined but may be refined after quality criteria are defined.
- Data extraction forms initially prepared during construction of the protocol will be amended when quality criteria are agreed.
- Data synthesis methods defined in the protocol may be amended once data has been collected.

The systematic reviews road map prepared by the Systematic Reviews Group at Berkley demonstrates the iterative nature of the systematic review process very clearly [15].

4. Planning

4.1 The need for a systematic review

The need for a systematic review arises from the requirement of researchers to summarise all existing information about some phenomenon in a thorough and unbiased manner. This may be in order to draw more general conclusion about some phenomenon than is possible from individual studies, or as a prelude to further research activities.

Prior to undertaking a systematic review, researchers should ensure that a systematic review is necessary. In particular, researchers should identify and review any existing systematic reviews of the phenomenon of interest against appropriate evaluation criteria. CRC [12] suggests the following checklist:

- What are the review's objectives?
- What sources were searched to identify primary studies? Were there any restrictions?
- What were the inclusion/exclusion criteria and how were they applied?
- What criteria were used to assess the quality of primary studies and how were they applied?
- How were the data extracted from the primary studies?
- How were the data synthesised? How were differences between studies investigated? How were the data combined? Was it reasonable to combine the studies? Do the conclusions flow from the evidence?

From a more general viewpoint, Greenlaugh [9] suggests the following questions:

- Can you find an important clinical question, which the review addressed? (Clearly, in software engineering, this should be adapted to refer to an important software engineering question.)
- Was a thorough search done of the appropriate databases and were other potentially important sources explored?
- Was methodological quality assessed and the trials weighted accordingly?
- How sensitive are the results to the way that the review has been done?
- Have numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

4.2 Development of a Review Protocol

A review protocol specifies the methods that will be used to undertake a specific systematic review. A pre-defined protocol is necessary to reduce the possibility researcher bias. For example, without a protocol, it is possible that the selection of individual studies or the analysis may be driven by researcher expectations. In medicine, review protocols are usually submitted to peer review.

The components of a protocol include all the elements of the review plus some additional planning information:

- Background. The rationale for the survey.
- The research questions that the review is intended answer.
- The strategy that will be used to search for primary studies including search terms and resources to be searched, resources include databases, specific journals, and conference proceedings. An initial scoping study can help determine an appropriate strategy.
- Study selection criteria and procedures. Study selection criteria determine criteria for including in, or excluding a study from, the systematic review. It is usually helpful to pilot the selection criteria on a subset of primary studies. The protocol should describe how the criteria will be applied e.g. how many assessors will evaluate each prospective primary study, and how disagreements among assessors will be resolved.

- Study quality assessment checklists and procedures. The researchers should develop quality checklists to assess the individual studies. The purpose of the quality assessment will guide the development of checklists.
- Data extraction strategy. This should define how the information required from each primary study would be obtained. If the data require manipulation or assumptions and inferences to be made, the protocol should specify an appropriate validation process.
- Synthesis of the extracted data. This should define the synthesis strategy. This should clarify whether or not a formal meta-analysis is intended and if so what techniques will be used.
- Project timetable. This should define the review plan.

4.2.1 The Research Question

4.2.1.1 Question Types

The most important activity during protocol is to formulate the research question. The Australian NHMR Guidelines [1] identify six types of health care questions that can be addressed by systematic reviews:

1. Assessing the effect of intervention.
2. Assessing the frequency or rate of a condition or disease.
3. Determining the performance of a diagnostic test.
4. Identifying aetiology and risk factors.
5. Identifying whether a condition can be predicted.
6. Assessing the economic value of an intervention or procedure.

In software engineering, it is not clear what the equivalent of a diagnostic test would be, but the other questions can be adapted to software engineering issues as follows:

- Assessing the effect of a software engineering technology.
- Assessing the frequency or rate of a project development factor such as the adoption of a technology, or the frequency or rate of project success or failure.
- Identifying cost and risk factors associated with a technology.
- Identifying the impact of technologies on reliability, performance and cost models.
- Cost benefit analysis of software technologies.

Medical guidelines often provide different guidelines and procedures for different types of question. This document does not go to this level of detail.

The critical issue in any systematic review is to ask the right question. In this context, the right question is usually one that:

- Is meaningful and important to practitioners as well as researchers. For example, researchers might be interested in whether a specific analysis technique leads to a significantly more accurate estimate of remaining defects after design inspections. However, a practitioner might want to know whether adopting a specific analysis technique to predict remaining defects is more effective than expert opinion at identifying design documents that require re-inspection.
- Will lead either to changes in current software engineering practice or to increased confidence in the value of current practice. For example, researchers

and practitioners would like to know under what conditions a project can safely adopt agile technologies and under what conditions it should not.
- Identify discrepancies between commonly held beliefs and reality.

Nonetheless, there are systematic reviews that ask questions that are primarily of interest to researchers. Such reviews ask questions that identify and/or scope future research activities. For example, a systematic review in a PhD thesis should identify the existing basis for the research student's work and make it clear where the proposed research fits into the current body of knowledge.

4.2.1.2 Question Structure

Medical guidelines recommend considering a question from three viewpoints:

- The population, i.e. the people affected by the intervention.
- The interventions usually a comparison between two or more alternative treatments.
- The outcomes, i.e. the clinical and economic factors that will be used to compare the interventions.

In addition, study designs appropriate to answering the review questions may be identified.

4.2.1.2.1 Population

In software engineering experiments, the populations might be any of the following:

- A specific software engineering role e.g. testers, managers.
- A type of software engineer, e.g. a novice or experienced engineer.
- An application area e.g. IT systems, command and control systems.

A question may refer to very specific population groups e.g. novice testers, or experienced software architects working on IT systems. In medicine the populations are defined in order to reduce the number of prospective primary studies. In software engineering far less primary studies are undertaken, thus, we may need to avoid any restriction on the population until we come to consider the practical implications of the systematic review.

4.2.1.2.2 Intervention

Interventions will be software technologies that address specific issues, for example, technologies to perform specific tasks such as requirements specification, system testing, or software cost estimation.

4.2.1.2.3 Outcomes

Outcomes should relate to factors of importance to practitioners such as improved reliability, reduced production costs, and reduced time to market. All relevant outcomes should be specified. For example, in some cases we require interventions that improve some aspect of software production without affecting another e.g. improved reliability with no increase in cost.

A particular problem for software engineering experiments is the use of surrogate measures for example, defects found during system testing as a surrogate for quality, or coupling measures for design quality. Studies that use surrogate measures may be misleading and conclusions based on such studies may be less robust.

4.2.1.2.4 Experimental designs

In medical studies, researches may be able to restrict systematic reviews to primary of studies of one particular type. For example, Cochrane reviews are usually restricted to randomised controlled trials (RCTs). In other circumstances, the nature of the question and the central issue being addressed may suggest that certain studies design are more appropriate than others. However, this approach can only be taken in a discipline where the amount of available research is a major problem. In software engineering, the paucity of primary studies is more likely to be the problem for systematic reviews and we are more likely to need protocols for aggregating information from studies of widely different types. A starting point for such aggregation is the ranking of primary studies of different types; this is discussed in Section 5.3.1.

4.2.2 Protocol Review

The protocol is a critical element of any systematic review. Researchers must agree a procedure for reviewing the protocol. If appropriate funding is available, a group of independent experts should be asked to review the protocol. The same experts can later be asked to review the final report.

PhD students should present their protocol to their supervisors for review and criticism.

5. Conducting the review

Once the protocol has been agreed, the review proper can start. This involves:
1. Identification of research
2. Selection of studies
3. Study quality assessment
4. Data extraction and monitoring progress
5. Data synthesis

Each of these stages will be discussed in this section. Although some stages must proceed sequentially, some stages can be undertaken simultaneously.

5.1 Identification of Research

The aim of a systematic review is to find as many primary studies relating to the research question as possible using an unbiased search strategy. For example, it is necessary to avoid language bias. The rigour of the search process is one factor that distinguishes systematic reviews from traditional reviews.

5.1.1 Generating a search strategy

It is necessary to determine and follow a search strategy. This should be developed in consultation with librarians. Search strategies are usually iterative and benefit from:

- Preliminary searches aimed at both identifying existing systematic reviews and assessing the volume of potentially relevant studies.

- Trial searchers using various combinations of search terms derived from the research question
- Reviews of research results
- Consultations with experts in the field

A general approach is to break down the question into individual facets i.e. population, intervention, outcomes, study designs. Then draw up a list of synonyms, abbreviations, and alternative spellings. Other terms can be obtained by considering subject headings used in journals and data bases. Sophisticated search strings can then be constructed using Boolean AND's and OR's.

Initial searches for primary studies can be undertaken initially using electronic databases but this is not sufficient. Other sources of evidence must also be searched (sometimes manually) including:

- Reference lists from relevant primary studies and review articles
- Journals (including company journals such as the IBM Journal of Research and Development), grey literature (i.e. technical reports, work in progress) and conference proceedings
- Research registers
- The Internet.

It is also important to identify specific researchers to approach directly for advice on appropriate source material.

Medical researchers have developed pre-packaged research strategies. Software Engineering Researchers need to develop and publish such strategies including identification of relevant electronic databases.

5.1.2 Publication Bias

Publication bias refers to the problem that *positive* results are more likely to be published than *negative* results. The concept of *positive* or *negative* results sometimes depends on the viewpoint of the researcher. (For example, evidence that full mastectomies were not always required for breast cancer was actually an extremely positive result for breast cancer sufferers). However, publication bias remains a problem particularly for formal experiments, where failure to reject the null hypothesis is considered less interesting than an experiment that is able to reject the null hypothesis.

Publication bias can lead to systematic bias in systematic reviews unless special efforts are made to address this problem. Many of the standard search strategies identified above are used to address this issue including:

- Scanning the grey literature
- Scanning conference proceedings
- Contacting experts and researches working in the area and asking them if they know of any unpublished results.

In addition, statistical analysis techniques can be used to identify the potential significance of publication bias (se Section 5.5.5).

5.1.3 Bibliography Management and Document Retrieval

Bibliographic packages such as Reference Manager or Endnote are very useful to manage the large number of reference that can be obtained from a thorough literature research.

Once reference lists have been finalised the full articles of potentially useful studies will need to be obtained. A logging system is needed to make sure all relevant studies are obtained.

5.1.4 Documenting the Search

The process of performing a systematic review must be transparent and replicable:

- The review must be documented in sufficient detail for readers to be able to assess the thoroughness of the search.
- The search should be documented as it occurs and changes noted and justified.
- The unfiltered search results should be saved and retained for possible reanalysis.

Procedures for documenting the search process are given in Table 1.

Table 1 Search process documentation

Data Source	Documentation
Electronic database	Name of database Search strategy for each database Date of search Years covered by search
Journal Hand Searches	Name of journal Years searched Any issues not searched
Conference proceedings	Title of proceedings Name of conference (if different) Title translation (if necessary) Journal name (if published as part of a journal)
Efforts to identify unpublished studies	Research groups and researchers contacted (Names and contact details) Research web sites searched (Date and URL)
Other sources	Date Searched/Contacted URL Any specific conditions pertaining to the search

5.2 Study Selection

Once the potentially relevant primary studies have been obtained, they need to be assessed for their actual relevance.

5.2.1 Study selection criteria

Study selection criteria are intended to identify those primary studies that provide direct evidence about the research question. In order to reduce the likelihood of bias, selection criteria should be decided during the protocol definition.

Inclusion and exclusion criteria should be based on the research question. They should be piloted to ensure that they can be reliably interpreted and that they classify studies correctly.

Issues:

- It is important to avoid, as far as possible, exclusions based on the language of the primary study. It is often possible to cope with French or German abstracts, but Japanese or Chinese papers are often difficult to access unless they have a well-structured English abstract.
- It is possible that inclusion decisions could be affected by knowledge of the authors, institutions, journals or year of publication. Some medical researchers have suggested reviews should be done after such information has been removed. However, it takes time to do this and experimental evidence suggests that masking the origin of primary studies does not improve reviews [3].

5.2.2 Study selection process

Study selection is a multistage process. Initially, selection criteria should be interpreted liberally, so that unless studies identified by the electronic and hand searchers can be clearly excluded based on titles and abstracts, full copies should be obtained.

Final inclusion/exclusion decisions should be made after the full texts have been retrieved. It is useful to maintain a list of excluded studies identifying the reason for exclusion.

5.2.3 Reliability of inclusion decisions

When two or more researchers assess each paper, agreement between researchers can be measured using the Cohen Kappa statistic [6]. Each disagreement must be discussed and resolved. This may be a matter of referring back to the protocol or may involve writing to the authors for additional information. Uncertainty about the inclusion/exclusion of some studies should be investigated by sensitivity analysis.

A single researcher should consider discussing included and excluded papers with an expert panel.

5.3 Study Quality Assessment

In addition, to general inclusion exclusion criteria, it is generally considered important to assess the “quality” of primary studies:

- To provide still more detailed inclusion/exclusion criteria.
- To investigate whether quality differences provide an explanation for differences in study results.
- As a means of weighting the importance of individual studies when results are being synthesised.
- To guide the interpretation of findings and determine the strength of inferences.
- To guide recommendations for further research.

An initial difficulty is that there is no agreed definition of study “quality”. However, the CRD Guidelines [12] and the Cochrane Reviewers’ Handbook [4] both suggest that quality relates to the extent to which the study minimises bias and maximises internal and external validity (see Table 2).

Table 2 Quality concept definitions

Term	Synonyms	Definition
Bias	Systematic error	A tendency to produce results that depart systematically from the 'true' results. Unbiased results are internally valid
Internal validity	Validity	The extent to which the design and conduct of the study are likely to prevent systematic error. Internal validity is a prerequisite for external validity.
External validity	Generalisability, Applicability	The extent to which the effects observed in the study are applicable outside of the study.

5.3.1 Quality Thresholds

The CRD Guideline [4] suggests using an assessment of study design to guarantee a minimum level of quality. The Australian National Health and Medical Research Council guidelines [2] suggest that study design is considered during assessment of evidence rather than during the appraisal and selection of studies. Both groups however suggest a hierarchy of study designs (see Table 3 and Table 4).

Table 3 CRD Hierarchy of evidence

Level	Description
1	Experimental studies (i.e. RCT with concealed allocation)
2	Quasi-experimental studies (i.e. studies without randomisation)
3	Controlled observational studies
3a	Cohort studies
3b	Case control studies
4	Observational studies without control groups
5	Expert opinion based on theory, laboratory research or consensus

Table 4 Australian NHMRC Study design hierarchy

Level I	Evidence obtained from a systematic review of all relevant randomised trials
Level II	Evidence obtained from at least one properly-designed randomised controlled trial
Level III-1	Evidence obtained from well-designed pseudo-randomised controlled trials (i.e. non-random allocation to treatment)
Level III-2	Evidence obtained from comparative studies with concurrent controls and allocation not randomised, cohort studies, case-control studies or interrupted time series with a control group.
Level III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group
Level IV	Evidence obtained from case series, either post-test or pretest/post-test

In order to understand Table 3 and Table 4 some additional definitions of studies types is given in Table 5, where *experimental studies* are those in which some conditions, particularly those concerning the allocation of participants to different treatment groups are under the control of investigator and *observational* studies are those in which uncontrolled variation in treatment or exposure among study participants is investigated.

Although the definitions given in Table 5 appear appropriate to software engineering studies (replacing the word disease with condition), it is important to note one critical difference between medical experiments and software engineering experiments. Most experiments performed in academic settings **cannot** be equated to randomised controlled trials (RCTs) in medicine.

Table 5 Definition of study designs

Design Type	Synonym	Basic Type	Definition	Source
Randomised Controlled Trial (RCT)	Randomised Clinical Trial	Experiment	An experiment in which investigators randomly allocate eligible people into intervention groups	[5]
Quasi-randomised trial	Pseudo-randomised controlled trial	Experiment	A study in which the allocation of participants to different intervention groups is controlled by the investigator but the method falls short of genuine randomisation and allocation concealment.	[12]
Cohort study	Follow-up study, incidence study, longitudinal study, prospective study	Observation	An observational study in which a defined group of people (the cohort) is followed over time. The outcomes of people in subsets are compared to examine for example people who were exposed to or not exposed (or exposed at different levels) to a particular intervention.	[12]
Concurrent cohort study		Observation	A study where a cohort is assembled in the present and followed into the future	[12]
Historical cohort study		Observation	A study where a cohort is identified from past records and followed from that time to the present.	[12]
Case-control study		Observation	Subjects with the outcome or disease and an appropriate group of controls without the outcome or disease are selected and information is obtained about the previous exposure to the treatment or other factor being studied	[2]
Historical control		Observation	Outcomes for a prospectively collected group of subjects exposed to a new treatment/intervention are compared with either a previously published series or previously treated subjects at the same institutions.	[2]
Interrupted time series		Observation	Trends in the outcomes or diseases are compared over multiple time points before and after introduction of the treatment/intervention or other factor being studied.	[2]
Cross-sectional study		Observation	Examination of relationships between diseases and other variables of interest as they exist in a defined population at one particular time	[12]
Case series		Observation	A group of subjects are exposed to the treatment or intervention	[2]
Post-test case series		Observation	A case series where only outcomes after the intervention are recorded in the case series, so no comparisons can be made.	[2]
Pre-test / post-test case series	Before-and-after study	Observation	A case series where outcomes are measured in subjects before and after exposure to the treatment/intervention for comparison.	[2]

RCTs involve real patients with real diseases receiving a new treatment to manage their condition. That is, RCTs are trials of treatment under its actual use conditions. The majority of academic experiments involve students doing constrained tasks in artificial environments. Thus, the major issue for software engineering study hierarchies is whether small-scale experiments are considered the equivalent of laboratory experiments and evaluated at the lowest level of evidence, or whether they should be ranked higher. In my opinion, they should be ranked higher than expert opinion. I would consider them equivalent in value to case series or observational studies without controls. Two other issues that need to be resolved are:

- Whether or not systematic reviews are included in the hierarchy.
- Whether or not expert opinion is included in the hierarchy.

The inclusion of systematic reviews depends on whether you are classifying individual studies or assessing the level of evidence. For assessing individual primary studies, systematic reviews are, of course, excluded. For assessing the level of evidence, systematic reviews should be considered the highest level of evidence. However, in contrast to the implication in the Australian Hierarchy in Table 4, I believe software engineers must consider systematic reviews of many types of primary study not only randomised controlled trials.

The Australian NHMRC guidelines [2] do not included expert opinion in their hierarchy. The authors remark that the exclusion is a result of studies identifying the fallibility of expert opinion. In software engineering we may have little empirical evidence, so may have to rely more on expert opinion than medical researchers. However, we need to recognise the weakness of such evidence.

Table 6 Study design hierarchy for Software Engineering

1	Evidence obtained from at least one properly-designed randomised controlled trial
2	Evidence obtained from well-designed pseudo-randomised controlled trials (i.e. non-random allocation to treatment)
3-1	Evidence obtained from comparative studies with concurrent controls and allocation not randomised, cohort studies, case-control studies or interrupted time series with a control group.
3-2	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group
4-1	Evidence obtained from a randomised experiment performed in an artificial setting
4-2	Evidence obtained from case series, either post-test or pre-test/post-test
4-3	Evidence obtained from a quasi-random experiment performed in an artificial setting
5	Evidence obtained from expert opinion based on theory or consensus

These considerations lead to the hierarchy shown in Table 6 for Software Engineering. Studies. This table includes reference to randomised controlled trials although I am aware of only one software engineering experiment that comes anywhere close to a randomised controlled trial in the sense that it undertakes an experiment in a real-life situation [11]. In this study, Jørgensen and Carelius requested a bid for a real project from a large number of commercial software companies in Norway. Companies were selected using stratified random sampling. Once the full sample was obtained, companies were randomly assigned to two groups. One group of companies were involved in pre-study phase and the bidding phase, the other companies were only involved in the bidding phase. The treatment in this case, was the pre-study activity, which involved companies providing an initial non-binding preliminary bid. One aspect that is not consistent with an RCT is that companies were paid for their time in order to compensate them for providing additional information to the experimenters. In addition, the study was not aimed at formal hypothesis testing, so the outcome was a possible explanatory theory rather than a statement of expected treatment effect.

Normally, primary study hierarchies are used to set a minimum requirement on the type of study included in the systematic review. In software engineering, we will usually accept all levels of evidence. The only threshold that might be viable would be to exclude level 5 evidence when there are a reasonable number of primary studies at a greater level (where a reasonable number must be decided by the researchers, but should be more than 2).

Categorising evidence hierarchies does not by itself solve the problem of how to accumulate evidence from studies in different categories. We discuss some fairly simple ideas in Section 5.5.4 used to present evidence, but we may need to identify new methods of accumulating evidence from different types of study. For example, Hardman and Ayton discuss a system to allow the accumulation of qualitative as well as quantitative evidence in the form arguments that are for or against proposition [13].

In addition, we need better understand the strength of evidence from different types of study. However, this is difficult. For example, there is no agreement among medical practitioners of the extent to which results from observational studies can really be trusted. Some medical researchers are critical of the reliance on RCTs and report cases where observational studies produced almost identical results to RCTs [8]. Concato and Horowitz suggest that improvements in reporting clinical conditions (i.e. collecting more information about individual patients and the reasons for assigning the patient to a particular treatment) would make observational studies as reliable as RCTs [7]. In contrast, Lawlor et al. discuss an example where results of a RCT proved that observational studies were incorrect [14]. Specifically, beneficial effects of vitamins in giving protection against heart disease found in two observational studies could not be detected in a randomised controlled trial. They suggest that better identification and adjustment for possible confounding factors would improve the reliability of observational studies. In addition, Vandenbroucke suggests that observational studies are appropriate for detecting negative side-effects, but not positive side-effects of treatments [17].

Observational studies and experiments in software engineering often have more in common with studies in the social sciences than medicine. For example, both social science and software engineering struggle with the problems both of defining and measuring constructs of interest, and of understanding the impact of experimental context on study results. From the viewpoint of social science, Shadish et al. provide a useful discussion of the study design and analysis methods that can improve the validity of experiments and quasi-experiments [16]. They emphasise the importance of identifying and either measuring or controlling confounding factors. They also discuss threats to validity across all elements of a study i.e. subjects, treatments, observations and settings.

5.3.2 Development of Quality Instruments

Once the primary studies have been selected a more detailed quality assessment needs to be made. This allows researchers to assess differences in the executions of studies within design categories. This information is important for data synthesis and interpretation of results. Detailed quality assessments are usually based on "quality instruments" which are checklists of factors that need to be assessed for each study. If quality items within a checklist are assigned numerical scales numerical assessments of quality can be obtained.

Checklists are usually derived from a consideration of factors that could bias study results. The CRD Guidelines [12], the Australian National Health and Medical Research Council Guidelines [1], and the Cochrane Reviewers' Handbook [4] all refer to four types of bias shown in Table 7. (I have amended the definitions (slightly) and protection mechanisms (considerably) to address software engineering rather than medicine.) In particular, medical researchers rely on "blinding" subjects and experimenters (i.e. making sure that neither the subject nor the researcher knows which treatment a subject is assigned to) to address performance and measurement bias. However, that protocol is often impossible for software engineering experiments.

Table 7 Types of Bias

Type	Synonyms	Definition	Protection mechanism
Selection bias	Allocation bias	Systematic difference between comparison groups with respect to treatment	Randomisation of a large number of subjects with concealment of the allocation method (e.g. allocation by computer program not experimenter choice).
Performance bias		Systematic difference is the conduct of comparison groups apart from the treatment being evaluated.	Replication of the studies using different experimenters. Use of experimenters with no personal interest in either treatment.
Measurement bias	Detection Bias	Systematic difference between the groups in how outcomes are ascertained.	Blinding outcome assessors to the treatments is sometimes possible.
Attrition bias	Exclusion bias	Systematic differences between comparison groups in terms of withdrawals or exclusions of participants from the study sample.	Reporting of the reasons for all withdrawals. Sensitivity analysis including all excluded participants.

The factors identified in Table 7 are refined into a quality instrument by considering:

- Generic items that relate to features of particular study designs such as lack of appropriate blinding, unreliable measurement techniques, inappropriate selection of subjects, and inappropriate statistical analysis.
- Specific items that relate to the review's subject area such as use of outcome measures inappropriate for answering the research question.

More detailed discussion of bias (or threats to validity) from the viewpoint of the social sciences rather than medicine can be found in Shadish et al. [16].

Examples of generic quality criteria for several types of study design are shown in Table 8. The items were derived from lists in [2] and [12].

If required, researchers may construct a measurement scale for each item. Whatever form the quality instrument takes, it should be assessed for reliability and usability in a pilot project before being applied to all the selected studies.

5.3.3 Using the Quality Instrument

Quality appraisal of each primary study allows researchers to group studies by quality prior to any synthesis of results. Researchers can then investigate whether there are systematic differences between primary studies in different quality groups.

Some researchers have suggested weighting results using quality scores. This idea is **not** recommended by any of the medical guidelines.

5.3.4 Limitations of Quality Assessment

Primary studies are often poorly reported, so it may not be possible to determine how to assess a quality criterion. It is possible to assume that because something wasn't reported, it wasn't done. This assumption may be incorrect. Researchers should attempt to obtain more information from the authors of the study.

Table 8 Example of Quality Criteria

Study type	Quality criteria
Cohort studies	How were subjects chosen for the new intervention?
	How were subjects selected for the comparison or control?
	Were drop-out rates and reasons for drop-out similar across intervention and unexposed groups?
	Does the study adequately control for demographic characteristics, and other potential confounding variables in the design or analysis?
	Was the measurement of outcomes unbiased (i.e. blinded to treatment group and comparable across groups)?
	Were there exclusions from the analysis?
Case-control studies	How were cases defined and selected?
	How were controls defined and selected? (I.e. were they randomly selected from the source population of the cases)
	How comparable are the cases and the controls with respect to potential confounding factors?
	Does the study adequately control for demographic characteristics, and other potential confounding variables in the design or analysis?
	Was measurement of the exposure to the factor of interest adequate and kept blinded to the case/control status?
	Were all selected subjects included in the analysis?
	Were interventions and other exposures assessed in the same way for cases and controls?
	Was an appropriate statistical analysis used (i.e. matched or unmatched)?
Case series	Is the study based on a representative sample from a relevant population?
	Are criteria for inclusion explicit?
	Were outcomes assessed using objective criteria?

There is limited evidence of relationships between factors that are thought to affect validity and actual study outcomes. Evidence suggests that inadequate concealment of allocation and lack of double-blinding result in over-estimates of treatment effects, but the impact of other quality factors is not supported by empirical evidence.

It is possible to identify inadequate or inappropriate statistical analysis, but without access to the original data it is not possible to correct the analysis. Very often software data is confidential and cannot therefore be made available to researchers. In some cases, software engineers may refuse to make their data available to other researchers because they want to continue publishing analyses of the data.

5.4 Data Extraction

The objective of this stage is to design data extraction forms to accurately record the information researchers obtain from the primary studies. To reduce the opportunity for bias, data extraction forms should be defined and piloted when the study protocol is defined.

5.4.1 Design of Data Extraction Forms

The data extraction forms must be designed to collect all the information needed to address the review questions and the study quality criteria. They must also collect all data items specified in the review synthesis strategy section of the protocol.

In most cases, data extraction will define a set of numerical values that should be extracted for each study (e.g. number of subjects, treatment effect, confidence intervals, etc.). Numerical data are important for any attempt to summarise the results of a set of primary studies and are a prerequisite for meta-analysis (i.e. statistical techniques aimed at integrating the results of the primary studies).

Data extraction forms need to be piloted on a sample of primary studies. If several researchers will use the forms, several researchers should take part in the pilot. The pilot studies are intended to assess both technical issues such as the completeness of the forms and usability issues such as the clarity of user instructions and the ordering of questions.

Electronic forms are useful and can facilitate subsequent analysis.

5.4.2 Contents of Data Collection Forms

In addition, to including all the questions needed to answer the review question and quality evaluation criteria, data collection forms should provide standard information including:

- Name of Review
- Date of Data extraction
- Title, authors, journal, publication details
- Space for additional notes

5.4.3 Data extraction procedures

Whenever feasible, data extraction should be performed independently by two or more researchers. Data from the researchers must be compared and disagreements resolved either by consensus among researchers or arbitration by an additional independent researcher. Uncertainties about any primary sources for which agreement cannot be reached should be investigated as part of any sensitively analyses. A separate form must be used to mark and correct errors or disagreements.

If several researchers each review different primary studies because time or resource constraints prevent all primary papers being assessed by at least two researchers, it is important to ensure employ some method of checking that researchers extract data in a consistent manner. For example, some papers should be reviewed by all researchers (e.g. a random sample of primary studies), so that inter-researcher consistency can be assessed.

For single researchers such as PhD students, other checking techniques must be used, for example supervisors should be asked to perform data extraction on a random sample of the primary studies and results cross-checked with those of the student.

5.4.4 Multiple publications of the same data

It is important to avoid including multiple publications of the same data in a systematic review synthesis because duplicate reports would seriously bias any results. It may be necessary to contact the authors to confirm whether or not reports refer to the same study. When there are duplicate publications, the most recent should be used.

5.4.5 Unpublished data, missing data and data requiring manipulation

If information is available from studies in progress, it should be included providing appropriate quality information about the study can be obtained and written permission is available from the researchers.

Reports do not always include all relevant data. They may also be poorly written and ambiguous. Again the authors should be contacted to obtain the required information.

Sometimes primary studies do not provide all the data but it is possible to recreate the required data by manipulating the published data. If any such manipulations are required, data should first be reported in the way they were reported. Data obtained by manipulation should be subject to sensitivity analysis.

5.5 Data Synthesis

Data synthesis involves collating and summarising the results of the included primary studies. Synthesis can be descriptive (non-quantitative). However, it is sometimes possible to complement a descriptive synthesis with a quantitative summary. Using statistical techniques to obtain a quantitative synthesis is referred to as *meta-analysis*. Description of meta-analysis methods is beyond the scope of this document, although techniques for displaying quantitative results will be described. (To learn more about meta-analysis see [4].)

The data synthesis activities should be specified in the review protocol. However, some issues cannot be resolved until the data is actually analysed, for example, subset analysis to investigate heterogeneity is not required if the results show no evidence of heterogeneity.

5.5.1 Descriptive synthesis

Extracted information about the studies (i.e. intervention, population, context, sample sizes, outcomes, study quality) should be tabulated in a manner consistent with the review question. Tables should be structured to highlight similarities and difference between study outcomes.

It is important to identify whether results from studies are consistent one with another (i.e. homogeneous) or inconsistent (e.g. heterogeneous). Results may be tabulated to display the impact of potential sources of heterogeneity, e.g. study type, study quality, and sample size.

Quantitative data should also be presented in tabular form including:

- Sample size for each intervention
- Estimates effect size for each intervention with standard errors for each effect
- Difference between the mean values for each intervention, and the confidence interval for the difference.
- Units used for measuring the effect.

5.5.2 Quantitative Synthesis

To synthesis quantitative results from different studies, study outcomes must be presented in a comparable way. Medical guidelines suggest different effect measures for different types of outcome.

Binary outcomes (Yes/No, Success/Failure) can be measured in several different ways:

- Odds. The ratio of the number of subjects in a group with an event to the number without an event. Thus if 20 projects in a group of 100 project failed to achieve budgetary targets, the odds would be 20/80 or 0.25.
- Risk (proportion, probability, rate) The proportion of subjects in a group observed to have an event. Thus, if 20 out of 100 projects failed to achieve budgetary targets, the risk would be 20/100 or 0.20.
- Odds ratio (OR). The ratio of the odds of an event in the experimental (or intervention) group to the odds of an event on the control group. An OR equal to one indicates no difference between the control and the intervention group. For undesirable outcomes a value less than one indicates that the intervention was successful in reducing risk, for a desirable outcome a value greater than one indicates that the intervention was successful in reducing risk.
- Relative risk (RR) (risk ratio, rate ratio). The ratio of risk in the intervention group to the risk in the control group. An RR of one indicates no difference between comparison groups. For undesirable events an RR less than one indicates the intervention was successful, for desirable events an RR greater than one indicates the intervention was successful.
- Absolute risk reduction (ARR) (risk difference, rate difference). The absolute difference in the event rate between the comparison groups. A difference of zero indicates no difference between the groups. For an undesirable outcome an ARR less than zero indicates a successful intervention, for a desirable outcome an ARR greater than zero indicates a successful intervention.

Each of these measures has advantages and disadvantages. For example, odds and odds ratios are criticised for not being well-understood by non-statisticians (other than gamblers), whereas risk measures are generally easier to understand. Alternatively statisticians prefer odd ratios because they have some mathematically desirable properties. Another issue is the relative measures are generally more consistent than absolute measures for statistical analysis, but decision makers need absolute values in order to assess the real benefit of an intervention.

Effect measures for continuous data include:

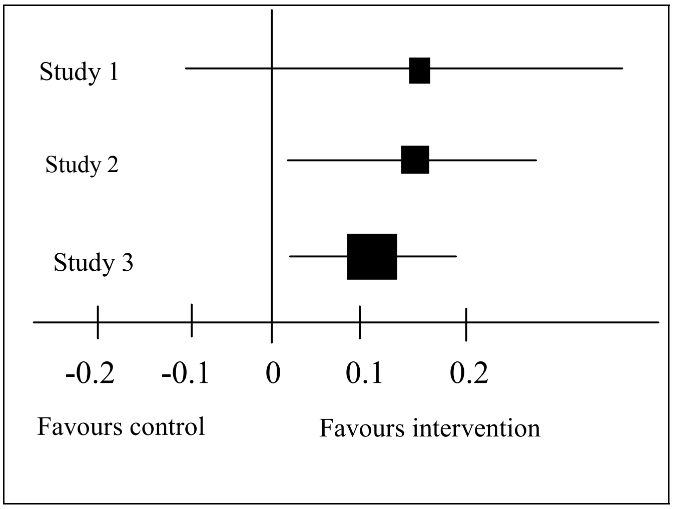
- Mean difference. The difference between the means of each group (control and intervention group).
- Weighted mean difference (WMD). When studies have measured the difference on the same scale, the weight give to each study is usually the inverse of the study variance
- Standardised mean difference (SMD). A common problem when summarising outcomes is that outcomes are often measured in different ways, for example, productivity might be measured in function points per hour, or lines of code per day. Quality might be measured as the probability of exhibiting one or more faults or the number of faults observed. When studies use different scales, the mean difference may be divided by an estimate of the within-groups standard deviation to produce a standardised value without any units. However, SMDs are only valid if the difference in the standard deviations reflect differences in the measurement scale, not real differences among trial populations.

5.5.3 Presentation of Quantitative Results

The most common mechanism for presenting quantitative results is a forest plot, as shown in Figure 1. A forest plot presents the means and variance for the difference for each study. The line represents the standard error of the difference, the box represents the mean difference and its size is proportional to the number of subjects in the study. A forest plot may also be annotated with the numerical information indicating the number of subjects in each group, the mean difference and the confidence interval on the mean. If a formal meta-analysis is undertaken, the bottom entry in a forest plot will be the summary estimate of the treatment difference and confidence interval for the summary difference.

Figure 1 represents the ideal result of a quantitative summary, the results of the studies basically agree. There is clearly a genuine treatment effect and a single overall summary statistics would be a good estimate of that effect. If effects were very different from study to study, our results would suggest heterogeneity. A single overall summary statistics would probably be of little value. The systematic review should continue with an investigation of the reasons for heterogeneity. To avoid the problems of post-hoc analysis, researchers should identify possible sources of heterogeneity when they construct the review protocol.

Figure 1 Example of a forest plot



5.5.4 Sensitivity analysis

Sensitivity analysis is much more important when a full meta-analysis is performed than when no formal meta-analysis is performed. Meta-analysis is used to provide an overall estimate of the treatment effect and its variability. In such cases, the results of the analysis should be repeated on various subsets of primary studies to determine whether the results are robust. The types of subsets selected would be:

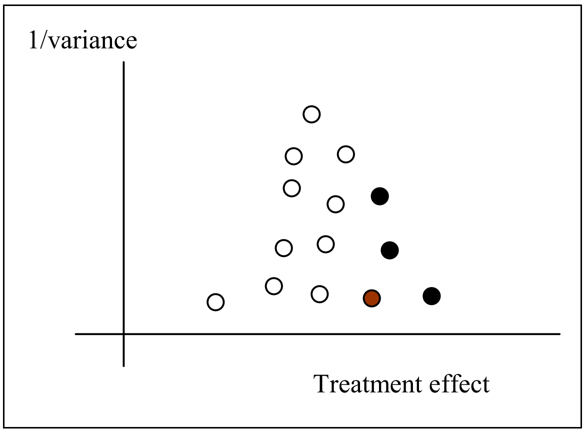
- High quality primary studies only.
- Primary studies of particular types.
- Primary studies for which data extraction presented no difficulties (i.e. excluding any studies where there was some residual disagreement about the data extracted).

When a formal meta-analysis is not undertaken, forest plots can be annotated to identify high quality primary studies, the studies can be presented in decreasing order of quality or in decreasing study type hierarchy order. Primary studies where there are queries about the data extracted can also be explicitly identified on the forest plot, by for example, using grey colouring for less reliability studies and black colouring for reliable studies.

5.5.5 Publication bias

Funnel plots are used to assess whether or not a systematic review is likely to be vulnerable to publication bias. Funnel plots plot the treatment effect (i.e. mean difference between intervention group and control) against the inverse of the variance or the sample size. A systematic review that exhibited the funnel shape shown in Figure 2 would be assumed **not** to be exhibiting evidence of publication bias. It would be consistent with studies based on small samples showing more variability in outcome than studies based on large samples. If, however, the points shown as filled-in black dots were not present, the plot would be asymmetric and it would suggest the presence of publication bias. This would suggest the results of the systematic survey must be treated with caution.

Figure 2 An example of a funnel plot



6. Reporting the review

It is important to communicate the results of a systematic review effectively. Usually systematic reviews will be reported in at least two formats:

1. In a technical report or in a section of a PhD thesis.
2. In a journal or conference paper.

A journal or conference paper will normally have a size restriction. In order to ensure that readers are able to properly evaluate the rigour and validity of a systematic review, journal papers should reference a technical report or thesis that contains all the details.

In addition, systematic reviews with important practical results may be summarised in non-technical articles in practitioner magazines, in press releases and in Web pages.

6.1 Structure for systematic review

The structure and contents of reports suggested in [12] is presented in Table 9. This structure is appropriate for technical reports and journals. For PhD theses, the entries marked with an asterisk are not likely to be relevant.

6.2 Peer Review

Journal articles will be peer reviewed as a matter of course. Experts review PhD theses as part of the examination process. In contrast, technical reports are not usually subjected to peer review. However, if systematic reviews are made available on the Web so that results are made available quickly to researchers and practitioners, it is worth organising a peer review. If an expert panel were assembled to review the study protocol, the same panel would be appropriate to undertake peer review of the systematic review report.

.

Table 9 Structure and contents of reports of systematic reviews

Section	Subsection	Scope	Comments
Title*			The title should be short but informative. It should be based on the question being asked. In journal papers, it should indicate that the study is a systematic review.
Authorship*			When research is done collaboratively, criteria for determining both who should be credited as an author, and the order of author's names should be defined in advance. The contribution of workers not credited as authors should be noted in the Acknowledgements section.
Executive summary or Structured Abstract*	Context	The importance of the research questions addressed by the review	A structured summary or abstract allows readers to assess quickly the relevance, quality and generality of a systematic review.
	Objectives	The questions addressed by the systematic review	
	Methods	Data Sources, Study selection, Quality Assessment and Data extraction	
	Results	Main finding including any meta-analysis results and sensitivity analyses.	
	Conclusions	Implications for practice and future research	
Background		Justification of the need for the review. Summary of previous reviews	Description of the software engineering technique being investigated and its potential importance
Review questions		Each review question should be specified	Identify primary and secondary review questions. Note this section may be included in the background section.
Review Methods	Data sources and search strategy		This should be based on the research protocol. Any changes to the original protocol should be reported.
	Study selection		
	Study quality assessment		
	Data extraction		
	Data synthesis		
Included and excluded studies		Inclusion and exclusion criteria List of excluded studies with rationale for exclusion	Study inclusion and exclusion criteria can sometimes best be represented as a flow diagram because studies will be excluded at different stages in the review for different reasons.

Results	Findings	Description of primary studies Results of any quantitative summaries Details of any meta-analysis	Non-quantitative summaries should be provided to summarise each of the studies and presented in tabular form. Quantitative summary results should be presented in tables and graphs
	Sensitivity analysis		
Discussion	Principal findings		These must correspond to the findings discussed in the results section
	Strengths and Weaknesses	Strength and weaknesses of the evidence included in the review Relation to other reviews, particularly considering any differences in quality and results.	A discussion of the validity of the evidence considering bias in the systematic review allows a reader to assess the reliance that may be placed on the collected evidence.
	Meaning of findings	Direction and magnitude of effect observed in summarised studies Applicability (generalisability) of the findings	Make clear to what extent the result imply causality by discussing the level of evidence. Discuss all benefits, adverse effects and risks. Discuss variations in effects and their reasons (for example are the treatment effects larger on larger projects).
Conclusions	Recommendations	Practical implications for software development	What are the implications of the results for practitioners?
		Unanswered questions and implications for future research	
Acknowledgements*		All persons who contributed to the research but did fulfil authorship criteria	
Conflict of Interest			Any secondary interest on the part of the researchers (e.g. a financial interest in the technology being evaluated) should be declared.
References and Appendices			Appendices can be used to list studies included and excluded from the study, to document search strategy details, and to list raw data from the included studies.

7. Final remarks

This report has presented a set of guidelines for planning conducting and reporting systematic review. The guidelines are based on guidelines used in medical research. However, it is important to recognise that software engineering research is not the same as medical research. We do not undertake randomised clinical trials, nor can we use blinding as a means to reduce distortions due to experimenter and subject expectations. Thus, software engineering research studies usually provide only weak evidence compared with RCTs.

We need to consider mechanisms to aggregate evidence from studies of different types and to understand the extent to which we can rely on such evidence. At present, these guidelines merely suggest that data from primary studies should be accompanied by information about the type of primary study and its quality. As yet, there is no definitive method for accumulating evidence from studies of different types. Furthermore, there is disagreement among medical researchers about how much reliance can be placed on evidence from studies other than RCTs. However, the limited number of primary studies in software engineering imply that it is critical to consider evidence from all types of primary study, including laboratory/academic experiments, and as well as evidence obtained from experts.

Finally, these guidelines are intended to assist PhD students as well as larger research groups. However, many of the steps in a systematic review assume that it will be undertaken by a large group of researchers. In the case of a single research (such as PhD student), we suggest the most important steps to undertake are:

- Developing a protocol.
- Defining the research question.
- Specifying what will be done to address the problem of a single researcher applying inclusion/exclusion criteria and undertaking all the data extraction.
- Defining the search strategy.
- Defining the data to be extracted from each primary study including quality data.
- Maintaining lists of included and excluded studies.
- Using the data synthesis guidelines.
- Using the reporting guidelines

8. References

[1] Australian National Health and Medical Research Council. How to review the evidence: systematic identification and review of the scientific literature, 2000. IBSN 186-4960329.

[2] Australian National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. February 2000, ISBN 0 642 43295 2.

[3] Berlin, J.A., Miles, C.G., Crigliano, M.D. Does blinding of readers affect the results of meta-analysis? Online J. Curr. Clin. Trials, 1997: Doc No 205.

[4] Cochrane Collaboration. Cochrane Reviewers' Handbook. Version 4.2.1. December 2003

[5] Cochrane Collaboration. The Cochrane Reviewers' Handbook Glossary, Version 4.1.5, December 2003.

[6] Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Pychol Bull (70) 1968, pp. 213-220.

[7] Concato, John and Horowitz, Ralph. I. Beyond randomised versus observational studies. The Lancet, vol363, Issue 9422, 22 May, 2004.

[8] Feinstein, A.R., and Horowitz, R.I. Problems with the "evidence" of "evidence-based medicine". Ann. J. Med., 1977, vol(103) pp529-535.

[9] Greenlaugh, Trisha. How to read a paper: Papers that summarise other papers (systematic reviews and meta-analysies. BMJ, 315, 1997, pp. 672-675.

[10] Jasperson, Jon (Sean), Butler, Brian S., Carte, Traci, A., Croes, Henry J.P., Saunders, Carol, S., and Zhemg, Weijun. Review: Power and Information Technology Research: A Metatriangulation Review. MIS Quarterly, 26(4): 397-459, December 2002.

[11] Jørgensen, Magne and Carelius, Gunnar J. An Empirical Study of Software Project Bidding, Submitted to IEEE TSE, 2004 (major revision required). http://www.simula.no/photo/desbidding16.pdf.

[12] Khan, Khalid, S., ter Riet, Gerben., Glanville, Julia., Sowden, Amanda, J. and Kleijnen, Jo. (eds) Undertaking Systematic Review of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition), NHS Centre for Reviews and Dissemination, University of York, IBSN 1 900640 20 1, March 2001.

[13] Hardman, David, K, and Ayton, Peter. Arguments for qualitative risk assessment: the StAR risk advisor. Expert Systems, Vol 14, No. 1., 1997, pp24-36.

[14] Lawlor, Debbie A., George Davey Smith, K Richard Bruckdorfer, Devi Kundu, Shah. Ebrahim Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? The Lancet, vol363, Issue 9422, 22 May, 2004.

[15] Pai, Madhukar., McCulloch, Michael., and Colford, Jack. Systematic Review: A Road Map Version 2.2. Systematic Reviews Group, UC Berkeley, 2002. [www.medepi.org/meta/guidelines/Berkeley_Systematic_Reviews_Road_Map_V2.2.pdf viewed 20 June 2004].

[16] Shadish, W.R., Cook, Thomas, D. and Campbell, Donald, T. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, 2002.

[17] Jan P Vandenbroucke. When are observational studies as credible as randomised trials? The Lancet, vol363, Issue 9422, 22 May, 2004.

Appendix 1 Steps in a systematic review

Guidelines for systematic review in the medical domain have different view of the process steps needed in a systematic review. The Systematic Reviews Group (UC Berkely) present a very detailed process model [15], other sources present a coarser process. These process steps are summarised in Table 10, which also attempts to collate the different processes.

Pai et al. [15] have specified the review process steps at a more detailed level of granularity than the other systematic review guidelines. In particular, they have made explicit the iterative nature of the start of a systematic review process. The start-up problem is not discussed in any of the other guidelines. However, it is clear that it is difficult to determine the review protocol without any idea as to the nature of the research question and vice-versa.

Table 10 Systematic review process proposed in different guidelines

Systematic Reviews Group ([15])	Australian National Health and Medical Research Council ([1])	Cochrane Reviewers Handbook ([4])	CRD Guidance ([12])
			Identification of the need for a review. Preparation of a proposal for a systematic review
Define the question & develop draft protocol Identify a few relevant studies and do a pilot study; specific inclusion/exclusion criteria, test forms and refine protocol.		Developing a protocol	Development of a review protocol
	Question Formulation	Formulating the problem	
Identify appropriate databases/sources. Run searches on all relevant data bases and sources. Save all citations (titles/abstracts) in a reference manager. Document search strategy.	Finding Studies	Locating and selecting studies for reviews	Identification of research Selection of studies
Researchers (at least 2) screen titles & abstracts. Researchers meet & resolve differences. Get full texts of all articles. Researchers do second screen. Articles remaining after second screen is the final set for inclusion			
Researchers extract data including quality data.	Appraisal and selection of studies	Assessment of study quality	Study quality assessment

Researchers meet to resolve disagreements on data Compute inter-rater reliability. Enter data into database management software		Collecting data	Data extraction & monitoring progress
Import data and analyse using meta-analysis software. Pool data if appropriate. Look for heterogeneity.	Summary and synthesis of relevant studies	Analysing & presenting results	Data synthesis
Interpret & present data. Discuss generalizability of conclusions and limitations of the review. Make recommendations for practice or policy, & research.	Determining the applicability of results. Reviewing and appraising the economics literature.	Interpreting the results	The report and recommendations. Getting evidence into practice.