The Methods for the Synthesis of Studies without Control Groups

Donna Fitzpatrick-Lewis • Donna Ciliska • Helen Thomas

June, 2009

National Collaborating Centre
for Methods and Tools

Centre de collaboration nationale
des méthodes et outils

The Methods for the Synthesis of Studies without Control Groups

Prepared for the National Collaborating Centre for Methods and Tools by

Donna Fitzpatrick-Lewis • Donna Ciliska • Helen Thomas

June, 2009

National Collaborating Centre for Methods and Tools (NCCMT)

School of Nursing, McMaster University

Suite 302, 1685 Main Street West

Hamilton, Ontario L8S 1G5

Telephone: (905) 525-9140, ext. 20455

Fax: (905 529-4184

Funded by the Public Health Agency of Canada

Affiliated with McMaster University

Production of this paper has been made possible through a financial contribution from the Public Health Agency of Canada. The views expressed herein do not necessarily represent the views of the Public Health Agency of Canada.

How to cite this resource:
Fitzpatrick-Lewis, D., Ciliska, D., & Thomas, H. (2009). *The Methods for the Synthesis of Studies without Control Groups*. Hamilton, ON: National Collaborating Centre for Methods and Tools. [http://www.nccmt.ca/pubs/non-RCT2_EN.pdf]

Contents

Key Recommendations:

1. Incorporate studies without control groups into systematic reviews, especially when there are no other studies to consider. Studies without control groups can also provide information on long-term effectiveness, rare events and adverse effects.
2. Do not generalize the direction of differences in outcome between randomized and non-controlled studies. While effect sizes are more often smaller in randomized controlled trials (RCTs), some show the reverse or the same results across study designs.
3. Use Whittemore and Knafl's (2005) approach to include results of non-randomized trials. Whittemore and Knafl provide a five-step integrative review method that incorporates problem identification, literature search, data evaluation, data analysis and presentation of data.
4. Recommendations for methods of systematic review including studies without control groups:
    a) Use a critical appraisal tool that has been tested for internal consistency, test-retest and inter-rater reliability, and at least face and criterion validity (Downs & Black, 1998; Slim, Nini, Forestier et al., 2003; Steuten, Vrijhoef, van-Merode et al., 2004; Thomas, Ciliska, Dobbins et al., 2004a; Zaza, Wright-De Aguero, Briss et al., 2000).
    b) Do not meta-analyse results from observational studies.
5. Suggestions to authors of primary studies:
    a) Ensure a strong and transparent study design in all non-randomized trials.
    b) Confirm that the research question determines the study design.
    c) Incorporate detailed information about the study; lack of information means that readers cannot determine what has or has not been done in the study design.
    d) Ensure consistency in terminology and descriptions across non-randomized studies to facilitate comparisons.
6. Implications for further research:
    a) Conduct reliability and validity testing of any critical appraisal tools.
    b) Conduct further studies to assess the different results obtained by randomized versus non-randomized trials.
7. Including data from non-randomized trials is challenging; however, several benefits can be achieved:
    a) A wider body of literature can be included in search strategies (Thomas, Harden, Oakley et al., 2004).
    b) The scope of the questions that can be answered may be broadened.
    c) This research can shed light on why, how and for whom interventions or programs succeed or fail (Oliver, Harden, Rees et al., 2005).

Summary

There is currently no standardized model for synthesizing results of studies that do not have control groups. The purpose of this paper is to examine synthesis methods, critical appraisal tools and studies that deal with appraising and synthesizing quantitative studies without control groups. RCTs are achievable in clinical settings, but public health interventions can rarely replicate the controlled environment of the clinic. Researchers, policy-makers and practice decision-makers often must rely on other types of study designs for their source of evidence. This paper will outline precautions to be aware of when including non-controlled studies as evidence, and recommends effective tools to use to analyse the results from these types of studies.

Methods

With the assistance of a librarian, five electronic databases were searched using keywords. In addition, reference lists of all relevant articles and grey literature were searched.

An article met the relevance criteria if:

- it was a primary study comparing results of the same intervention using randomized controlled trials (RCTs) versus designs without control groups;
- it was a synthesis of methods used to review studies without control groups;
- at least one of these designs was included: one group cohort (before/after or just after), cross-sectional, interrupted time series or mixed methods;
- it was published between 1995 and November 2007.

A paper was excluded if:

- the focus was RCT, clinical controlled trials or quasi-randomized studies;
- the focus was a cohort analytic study (before/after on two or more groups);
- the study was qualitative only;
- it was published before 1995.

Thirty-six papers passed the relevance test and full data extraction was completed on those papers. Many of the included studies were observational studies. It is important to note that while there are many definitions of observational studies, only those without control groups were included in this paper.

Findings

Fourteen articles provided critical appraisal tools that could assist in judging the methodological rigour of primary studies. Primary studies for a systematic review should include a quality assessment of the studies. A critical appraisal tool is essential when sifting through evidence in primary studies without control groups. These tools can provide information about the strengths and weaknesses of information. Critical appraisal tools assist the reader in assessing the methodological quality, managing confounders, determining potential biases and performing data analysis. These tools are particularly helpful where the criteria for

assessing RCTs are not applicable. Many of the tools have not been tested for validity and reliability; this testing might be a sensible next step.

Eleven systematic reviews incorporated primary studies without control groups, or examined critical appraisal tools for quality assessment of studies without control groups. The reviews that examined the feasibility of combining the results of RCTs and non-RCTs suggest that well-designed studies can produce similar results, regardless of the study type. Many of these reviews suggest that poor execution of the primary study design result in lower quality scoring, rather than the study design itself. They cited under-reporting of methodology, poor management of confounders and/or lack of consistent analytical terminology as problematic when trying to compare data between studies.

Six papers examined various methods of synthesizing and integrating quantitative data from studies without control groups. The authors agreed that there is benefit to incorporating evidence from diverse study designs, but quality assessment is paramount for providing trustworthy evidence. Primary studies that are clear about potential bias and probable impact on effect can be usefully incorporated into systematic reviews.

Three articles described methods of synthesizing data from both qualitative and quantitative studies. Incorporating qualitative and quantitative data in one report can help build an understanding of the success or failure of interventions. Assessing the quality of studies without control groups can be problematic, especially if quality is assessed with traditional criteria used in systematic reviews or meta-analyses. Results of quality assessment were not used as exclusion criteria, but could be incorporated in the discussion of findings. Grouping findings according to a thematic and aggregate method was offered as an appropriate and alternative approach to meta-analysis. Researchers could also incorporate some of the processes used in systematic reviews, such as having two people review and perform data extraction. Most of these authors suggested a narrative approach for presenting results when meta-analysis was not appropriate.

Two papers compared the results of different methodologies applied to the same intervention. Those articles pointed out that it is useful to consider different designs in reviews. Not only may they be the best available information, but they also can be more able to answer questions related to long-term effectiveness, adverse events and rare events. The strengths of including different designs in reviews outweigh the limitations.

Recommendations

1. Incorporate studies without control groups into systematic reviews, especially when there are no other studies to consider. Studies without control groups can also provide information on long-term effectiveness, rare events and adverse effects.
2. Do not generalize the direction of differences in outcome between randomized and non-controlled studies. While effect sizes are more often smaller in randomized controlled trials (RCTs), some show the reverse or the same results across study designs.
3. Use Whittemore and Knafl's (2005) approach to include results of non-randomized trials. Whittemore and Knafl provide a five-step integrative review method that

incorporates problem identification, literature search, data evaluation, data analysis and presentation of data.

4. Recommendations for methods of systematic review including studies without control groups:
    a) Use a critical appraisal tool that has been tested for internal consistency, test-retest and inter-rater reliability, and at least face and criterion validity (Downs & Black, 1998; Slim, Nini, Forestier et al., 2003; Steuten, Vrijhoef, van-Merode et al., 2004; Thomas, Ciliska, Dobbins et al., 2004a; Zaza, Wright-De Aguero, Briss et al., 2000).
    b) Do not meta-analyse results from observational studies.
5. Suggestions to authors of primary studies:
    a) Ensure a strong and transparent study design in all non-randomized trials.
    b) Confirm that the research question determines the study design.
    c) Incorporate detailed information about the study; lack of information means that readers cannot determine what has or has not been done in the study design.
    d) Ensure consistency in terminology and descriptions across non-randomized studies to facilitate comparisons.
6. Implications for further research:
    a) Conduct reliability and validity testing of any critical appraisal tools.
    b) Conduct further studies to assess the different results obtained by randomized versus non-randomized trials.
7. Including data from non-randomized trials is challenging; however, several benefits can be achieved:
    a) A wider body of literature can be included in search strategies (Thomas, Harden, Oakley et al., 2004).
    b) The scope of the questions that can be answered may be broadened.
    c) This research can shed light on why, how and for whom interventions or programs succeed or fail (Oliver, Harden, Rees et al., 2005).

Introduction

The overriding goal of public health policy and practice is to protect people's health. To achieve this goal, public health professionals consider evidence when making decisions. Historically, scientists have preferred the randomized controlled trial (RCT) study design to provide the least biased results when analysing the effectiveness of interventions. An RCT may provide rigorous study results for effectiveness questions; however, it might not be the best study design for other types of research questions (DiCenso, Prevost, Benefield et al., 2004). In public health and health promotion, it may be difficult to use randomized trials for a particular question; does that mean we should ignore evidence from other methods? Some researchers believe that only RCTs have the scientific rigour to produce valuable evidence.

In clinical settings, RCTs are achievable. But public health interventions can rarely replicate the controlled environment of the clinic. These interventions are often community-based and recruitment can be challenging, creating a selection bias. As a result, maintaining pure control groups without cross-contamination may be impossible or impractical. Public health researchers must often rely on other types of study designs, often classified as lower on the "hierarchy of evidence" (Ogilvie, Egan, Hamilton et al., 2005). However, observational studies, for example, can provide valuable information for public health decision-making. Qualitative information can supplement statistical data by helping us understand the 'who, why and how' of intervention success or failure. There is a whole body of literature about the methods for synthesizing qualitative studies (meta-synthesis), but these methods are beyond the scope of this paper.

Due to the greater potential for bias in observational studies, they are often excluded from systematic reviews of treatments. However, they offer several advantages to RCTs, including: being less costly; allowing for larger sample sizes; and providing more long-term outcome measurements (Benson & Hartz, 2000). As well, observational studies may provide the "best available evidence" in certain situations (Ogilvie et al., 2005).

There is currently no standardized model for synthesizing the results of studies that do not have control groups. The purpose of this paper is to examine synthesis methods, critical appraisal tools and studies that deal with appraising and synthesizing the results from quantitative studies that lack control groups. This paper will also outline some precautions to take when using non-controlled studies as evidence, and will make recommendations for processes and tools.

Methods

The methods used to collect the literature for this paper included:

- a comprehensive literature search of published literature from January 1995 to October 2007;
- review and retrieval of references from relevant articles;
- a search and retrieval of potentially relevant grey literature.

The initial search was conducted by a skilled librarian, using the following key search words:

review, systematic review, literature review, meta-analysis, not random and not randomized clinical trials. These key word searches were conducted in a several databases: PsycINFO, OVID MEDLINE, EMBASE, CINAHL and Scholars Portal. The initial search located 9324 potentially relevant articles. Titles and abstracts were read independently by two reviewers who passed on 182 articles for full relevance testing.

Relevance

An article was included for full relevance testing if:

- it was a primary study comparing results of the same intervention using randomized trials versus designs without control groups;
- it was a synthesis of methods used to review studies without control groups;
- at least one of these designs was included: one group cohort (before/after or just after), cross-sectional, interrupted time series or mixed methods;
- it was published between 1995 and November 2007.

A paper was excluded if:

- the focus was RCT, clinical controlled trials or quasi-randomized studies;
- the focus was a cohort analytic study (before/after on two or more groups);
- the study was qualitative only;
- it was published before 1995.

Thirty-six papers passed the relevance test, and researchers completed full data extraction on those papers. Many of the final papers were observational studies. There are many definitions for observational studies; for this paper, only observational studies with no control groups were considered relevant.

Data Synthesis

Researchers synthesized 36 papers in a narrative format by topic:

- critical appraisal tools
- systematic reviews
- methods for synthesizing quantitative studies without control groups
- combined qualitative and quantitative studies
- comparing results using differing methods

Some papers could have been included in more than one section; however, we placed the papers in the groups that appeared most suitable.

The paper provides a narrative synopsis of the relevant papers. The results are also detailed in table format (See Appendix 1).

Findings

Critical Appraisal Tools

A total of 14 papers described critical appraisal tools.

Rangel, Kelsey, Colby et al., (2003) addressed the lack of a standardized measurement tool to determine the quality of observational studies. They developed a quality assessment instrument for non-randomized studies that incorporated 30 items within three subscales. The subscales assessed clinical relevance, reporting methodology and the strength of conclusions. Global ratings were determined by combining each subscale, and the studies were thereby rated as 'good,' 'fair' or 'poor.' They achieved high inter-rater reliability with levels of agreement reaching 84.6% concordance of results (n = 1573 items). For all individual subscales, there was range of 73.3% to 85.8% (n = 60 to 1258). The potential applications for this tool include:

- facilitating the development of standardized methodology for conducting systematic reviews of retrospective data;
- helping to develop a strategy for meta-analysis of non-randomized trials;
- developing standardized reporting guidelines for use in peer-reviewed journals.

Margetts et al. (1995) developed a scoring system to judge the scientific quality of observational epidemiological studies linking diet with the risk of cancer. Case-control and cohort studies were reviewed separately because some of the reliability markers applied to only one of these study types. Case-control studies were rated on three broad areas: quality of dietary assessment; recruitment of subject; and the analysis of results. Cohort studies were scored on four areas: dietary assessment, definition of cohort, ascertainment of disease cases and analysis of results. Inter-rater reliability was assessed and resulted in a high level of agreement between reviewers, with the cohort reviewers attaining a slightly higher level of agreement. This prototype scoring system helped the authors to describe the epidemiological data on diet and cancer; however it may not be generalizable to other topics.

Downs and Black (1998) developed and tested a checklist to assess randomized and non-randomized studies. They used epidemiological principles, reviews of study designs and existing checklists to develop their checklist, consisting of 26 items distributed between five subscales:

- reporting
- external validity
- bias
- confounding
- power

The maximum score was 31. Twenty-three of the questions could be asked of any study of any health care intervention. The other three questions were topic sensitive and were customized to provide the raters with information on confounders, main outcomes and the sam-

ple size required for clinical and statistical significance ($p<0.05$). The quality index had high internal consistency, good test-retest and inter-rater reliability and good face and criterion validity. It performed as well as other established checklists on randomized trials, and there was little difference between its performance with non-randomized and randomized studies.

To address the lack of validated instruments to measure quality of observational or non-randomized trials, Slim et al. (2003) tested a methodological index for non-randomized studies (MINORS). This tool includes 12 items (the first eight items were designed specifically for non-comparison trials):

- a clearly stated aim
- inclusion of consecutive patients
- prospective collection of data
- endpoints appropriate to the aim of the study
- unbiased assessment of the study endpoints
- a follow-up period appropriate to the aim of the study
- loss to follow-up less than 5%
- prospective calculation of the study size
- an adequate control group
- contemporary groups
- baseline equivalence of groups
- adequate statistical analysis

This tool had good reliability, internal consistency and validity.

Zaza et al. (2000) developed a format to classify and provide descriptive components of public health inventions and assess the quality of the study's execution. The authors acknowledged that all study designs have issues particular to their specific design, therefore the quality questions were developed to reflect those specific issues. For example, the authors suggested that while blinding is an important component to consider for randomized trials, it is not appropriate to assess validity in a time series design. Zaza et al. provided categories of questions to assess potential threats to the validity of results in primary studies, including:

- description of the invention
- sampling
- reliability and validity of outcome measurements
- the appropriateness of the statistical analysis
- the interpretation of results

One aim of their approach was to design a tool that was flexible enough to be useful in the assessment of divergent study designs and interventions. The standardized abstraction form and procedure improved the validity and reliability of the Guide to Community Prevention Services: Systematic Reviews and Evidence-Based Recommendations.

Greer et al. (2000) presented and discussed a system of grading evidence that was developed by the Institute for Clinical Systems Improvement (ICSI). This approach was developed in response to feedback indicating that an existing system was not working. Its goal is to assist busy practitioners who need to judge the quality of evidence. After reviewing many other grading systems, the ICSI working group agreed that it was important to separate the evaluation of the individual research reports from the assessment of the totality of the evidence supporting the conclusion. The subsequent worksheet for grading evidence—the primary tool—included a statement of conclusion, a summary of the research reports that support or dispute the conclusion, assignment of classes and quality markers to the research reports and assignment of a grade to the conclusion. Primary reports of new data collection were assigned a letter grade of A to D, depending on the type of study: randomized controlled trial; cohort study; non-randomized trial with concurrent or historic controls; case-control study; study of sensitivity and specificity of a diagnostic test; population-based descriptive study; cross-sectional study; case series; or case report. The report also included a five-point quality system to rate both positive and negative study design attributes, with ratings of plus, negative or neutral. The questions included:

- a priori inclusion/exclusion criteria
- bias
- statistical or clinically significant results
- generalizability of results to other populations
- clearly outlined study design

This system has been applied to more than 40 ICSI guidelines and technology assessment reports. The authors reported that the system appears to be successful in reducing the complexity of other grading systems, while yielding defendable classification of conclusions based on the strength of the underlying evidence.

Steuten et al. (2004) critiqued the tools used to assess the methodological quality of the Health Technology Assessment (HTA) of disease management. The authors developed an inventory of problems that arise when assessing HTA studies, including: study design (RCT vs. observational); criteria for selection and restriction of patients; baseline and outcome measures; blinding of patients and providers; the description of complex, multifaceted interventions; and avoidance of co-interventions. They developed, proposed and validated a new instrument for assessing the methodological quality of HTA for disease management that includes four components:

- study population
- description of intervention
- measurement of outcomes
- data analysis/presentation of data

The instrument includes 15 items that cover internal validity, external validity and statistical considerations. This instrument is reliable in terms of hierarchical ranking, test-retest reliability and inter-rater reliability when applied to HTAs of disease management.

Two reports (Chou & Helfand, 2005; McIntosh, Woolacott, & Bagnall, 2004) suggested

that evidence of harm is as important as evidence of effect. However, standard systematic reviews focusing on effectiveness and randomized trials may not be an efficient model for evaluating harm. Literature related to harm is found in unpublished trials, observational studies and grey literature. Applying existing quality checklists to this data set was problematic and did not readily capture the types of information pertinent to harmful effects. McIntosh et al. adapted published checklists to reflect information about harmful effects found in observational cohort studies. This new checklist also incorporated items such as how and when events were reported, and whether the time at which they occurred during the study was recorded. However, the greatest barrier to applying any checklist was the lack of methodological detail provided by the primary study authors. The researchers also included narrative outcome reports. Chou and Helfand (2005) examined case reports and observational studies of harmful effects of treatment. They too found quality assessment difficult, as instruments for assessing these types of studies were rare or inconsistent. Developmental rigour, scope, and the number and type of items used were all contentious issues. They pilot-tested a quality assessment tool for RCTs, clinical control trials (CCTs) and cohort studies for an uncontrolled surgical series that reported complications from carotid endarterectomy. This eight-point tool provided a ranking of 'good,' 'fair' or 'poor.' The tool was tested on studies for one intervention. The authors cautioned researchers to avoid inappropriate pooling of statistical results from observational studies, given the potential for bias due to inadequate control of confounders and selection bias.

The GRADE Working Group (Atkins, Briss, Eccles et al., 2005) reported on the development and pilot testing of the GRADE approach to assessing evidence and recommendations. Twelve evidence profiles were independently graded by 17 judges who all had experience using other approaches to assessing evidence and recommendations. Each evidence profile was based on information available in a systematic review and included two tables, one for quality assessment and one for the summary of the findings. The quality assessment process was designed so that each outcome was evaluated separately. The outcome table displayed the number of studies that had reported that outcome, the study design (RCTs or observational) and the quality of the studies. Assessment of the quality of outcome revealed that factors other than study design, quality, consistency and directness affected the reviewer's judgment about quality. These other factors included: sparse data, strong associations, publication bias, dose response and situations where all plausible confounders strengthened rather than weakened confidence in the direction of effect. The judges were asked about the ease of understanding and the sensibility of the approach; they generally agreed that the GRADE approach was clear, understandable and sensible.

Khan et al. (2001) provided information on what needs to be included in any quality assessment checklist. In effectiveness trials, RCTs should be considered first when they are available and well-designed. In the absence of good RCTs, well-designed quasi-experimental and observational studies should be considered. Khan et al. provided several questions to consider when assessing the quality of observational studies. They cautioned that although conclusions can be drawn from these studies, it may be unclear whether the groups within studies are comparable. The authors suggested that results need to be reported with caution, and often with recommendations for further research.

Thomas, Ciliska, Dobbins and Micucci (2004) used a quality assessment instrument that

allowed for the methodological rating of studies that were not randomized control trials, for inclusion in a systematic review. This method was specifically developed for quality assessment of public health research. Quality assessment components included:

- selection bias
- study design
- confounders
- blinding
- data collection methods
- withdrawals or dropouts

Studies rated with this system achieved a ‘strong,’ ‘moderate’ or ‘weak’ methodological rating. The instrument had good inter-rater and test-retest reliability. It also had proven content and criterion validity, randomized control trials and clinical control trials were rated as strong and non-RCTs such as cohorts and observational studies were rated as moderate. However, non-randomized studies could achieve an overall study rating of ‘moderate’ if the other components of the study were strong. Data is extracted from studies and reported in a narrative format. Meta-analysis was rarely done as it was found that public health study populations and interventions are often too heterogeneous for this type of analysis. Data extraction allowed for reporting statistically significant (or lack of) effects from each study, and the researchers commented on the clinical meaningfulness of any statistically significant effect.

Ramsay et al. (2003) reviewed the quality of Interrupted Time Series (ITS) using studies included in two systematic reviews. They developed a quality assessment tool that incorporated the following criteria:

- the intervention occurred independently of other changes over time;
- the intervention was unlikely to affect data collection;
- the outcome was assessed blindly or measured objectively;
- the outcome was reliable or measured objectively;
- the composition of the data set at each time point covered at least 80% of all participants in the study;
- the shape of the intervention effect was pre-specified;
- the rationale for the number and spacing of data points was described;
- the study was analysed appropriately using time series techniques.

They found that 37 of the 58 studies were not analysed appropriately, and that 33 of those should be re-analysed. Poor study designs, insufficient power and inappropriate analysis of ITS studies resulted in misleading conclusions being presented in the published literature. The authors recommend that all data from ITS studies be re-analysed before inclusion in reviews.

Conn and Rantz (2003) explored ways in which researchers can manage quality when considering primary studies within a meta-analysis. While there are more than 100 scales available to measure the quality of these studies, they vary in size, composition, complexity and extent of development. Establishing the validity of the scales is complicated and challenging. Scales are not reliable and accurate for all areas of science, and application of the

scales can be problematic and inconsistent. Conn and Rantz outlined three methods that could potentially manage quality:

- setting a quality threshold
- weighting by quality scoring
- considering quality as an empirical question

However, as a stand-alone method, each has limitations. For instance, setting a quality threshold may result in the exclusion of less rigorous research—data that may contain important information. Weighting by study quality can be limited by the scaling tools used and some potentially inherent problems such as inter-rater reliability. Finally, considering quality as an empirical question may lead to overall measures masking interesting effects of individual components of quality on effect-size estimates. To accommodate for these limitations, researchers may need to combine strategies to include studies that are rated as methodologically weak but still contain important information.

Section Summary

Reviewing literature to be included in a systematic review should include a quality assessment. A critical appraisal tool is essential when sifting through evidence in primary studies without control groups. These tools can provide information regarding the strengths and weaknesses of information. Critical appraisal tools assist the reader in assessing the evidence provided based on methodological quality, management of confounders, potential biases and data analysis. These tools are particularly helpful where systems of assessment used for RCTs are not applicable. Many of the tools presented here have not been tested for validity and reliability. This is a reasonable next step.

Systematic Reviews

Eleven systematic reviews considered the question of synthesizing data from studies without control groups.

Linde, Scholz, Melchart and Willich (2002) examined whether systematic reviews should include both randomized and non-randomized studies. The researchers examined the data on the use of acupuncture for chronic headaches to explore the following questions:

1. Do randomized and non-randomized studies of acupuncture for chronic headaches differ in regard to patients, interventions, design-independent quality aspects and response rates?
2. Do non-randomized studies provide relevant additional information (results of long-term outcomes, prognostic factors, adverse effects or complications and response rates in representative and well-defined groups of patients)?
3. If response rates in randomized and non-randomized patients differ, what are the possible explanations?

Fifty-nine studies met the inclusion criteria, of which 24 were randomized trials and 35 were non-randomized trials (five non-randomized control trials, 10 prospective uncontrolled stud-

ies, 10 case series and 10 cross-sectional surveys). A total of 535 patients received acupuncture treatment in the 24 RCTs, compared with 2695 in the 35 non-randomized studies. On average, randomized trials had smaller sample sizes, met more quality criteria and had lower response rates to treatment (0.59; 0.40–0.69) vs. (0.78; 0.72–0.83). Regardless of randomization status, studies meeting more quality criteria had lower response to treatment rates. Follow-up time was not significantly greater in the non-randomized studies. In the case of acupuncture for chronic headaches, non-randomized studies confirmed the findings in a previous study (Greenhalgh, Robert, Macfarlane et al., 2005) that the treatment was likely to be effective. However, non-randomized studies provided little additional relevant information on long-term effects, prognostic factors or adverse effects. In general, the authors concluded that non-randomized studies of good quality yielded results similar to RCTs. This suggests that their inclusion in systematic reviews, while increasing the workload for authors of these reviews, may provide a useful extension for generalizability.

MacLehose et al. (2000) investigated the association between methodological quality and estimates of effectiveness by comparing RCTs and quasi-experimental and observational (QEO) studies. The authors employed two strategies when reviewing the literature:

1. comparing RCT and QEO study estimates of effectiveness of any intervention, where both estimates were reported in one paper;
2. comparing the RCT and QEO estimates of effectiveness for specified interventions, where the estimates were reported in different papers.

For strategy 1, the authors identified 14 papers containing 38 comparisons, of which 13 were classified as high quality and 25 were classified as low. Quality assessment criteria included:

- study estimates derived from the same population;
- blinding of outcome assessors;
- the extent to which the QEO study estimate took possible confounding into account.

The findings indicated that the discrepancies between RCT and QEO study estimates of effect size and outcome frequency for intervention and control groups were smaller for high-quality than for low-quality comparisons. In addition, there was no tendency for QEO study estimates of effect size to be more extreme for high-quality comparisons than low-quality RCTs. For strategy 2, the specific interventions included mammography screening to reduce breast cancer mortality and folic acid supplements to reduce neutral tube defects. The authors identified 34 papers, of which 12 papers were RCTs, 11 were non-randomized trials or cohort studies and nine were matched or unmatched case-control studies. These studies were assessed based on

- the quality of the reporting;
- the generalizability of the results;
- the extent to which estimates of effectiveness may have been subject to bias or confounding.

Cohort and case-control studies had lower quality scores than RCTs, and cohort studies had

lower quality scores than case-control studies. Meta-regression of study attributes against relative risk estimates showed no association between effect size and study quality. Estimates for RCTs and cohort studies were not significantly different; however, case-control studies gave significantly different estimates in the mammography studies (greater benefit) and the folic acid studies (less benefit).

Britton et al. (1998) explored issues related to the process of randomization that may affect the validity of conclusions drawn from the results of RCTs and non-randomized studies (including studies without control groups). The authors asked four research questions to examine the issues of randomization and validity:

1. Do non-randomized studies differ systemically from RCTs in terms of treatment effect?
2. Are there systematic differences between the included and excluded individuals, and do these influence the measured treatment effect?
3. To what extent is it possible to adjust for baseline differences between study groups?
4. How important is patient preference in terms of outcomes?

Their study included 18 papers that directly compared the results from RCTs and non-randomized studies. The results found no consistent reporting of larger or smaller estimates of treatment effect based on study design. The authors also reported that the intervention type did not seem to be influential; however, they cautioned that more study is needed to validate that inference. In RCTs, blanket exclusions were common, and the number of included eligible subjects ranged from 1% to 100%. Large clinical databases containing detailed information of patient severity and prognosis were used instead of RCTs. Where the database subjects were selected according to the same inclusion criteria as RCTs, the treatment effects of the two designs were similar. Documentation of the characteristics of eligible individuals who did not participate in the trials was poorly recorded in most RCTs. Treatment effect measures in RCTs may be exaggerated due to the larger participation of university and teaching centres as compared to non-randomized studies. In non-randomized studies, adjustment for differences often changed the treatment effect size, but not significantly; more importantly, the direction of the change was consistent. Only four papers addressed the role of patient preference on results. In those papers, preference accounted for some of the observed differences between study designs. Britton et al. concluded:

- a well-designed non-randomized study is preferable to a small, poorly designed and exclusive RCT;
- RCTs should include a wide range of practice settings;
- study populations should be representative of all patients receiving the intervention;
- exclusions for administrative convenience should be rejected;
- differences should be minimized by ensuring that subjects in both kinds of study are comparable.

Lemmer, Grellier and Stevens (1999) modified the Cochrane Collaboration Protocol for Systematic Reviews for a report that explored the evidence for decision-making and health visits (home health care provision) in Britain. Health visiting and decision-making are influenced

by factors such as social and environmental forces, which tend not to be captured in the evidence emerging from randomized control trials. The adapted Cochrane Protocol omitted sections that were felt to be inappropriate for non-RCT research papers. The methodological rigour and relevance of each article was determined with a numerical score ranging from 8 (maximum) to 1 (minimum). The authors did not provide the variables used to measure methodological rigour. Two members of the research team reviewed each article, and weekly consensus meetings were held to reach agreement on scoring differences. The scores were intended to provide a rapid overview of the methodology and relevance of each article; however, the final aggregate score provided no indication of the allocation of marks. Reviewers commented on the appropriateness and significance of each article, but these comments were not part of the scoring. The authors found that the scoring system was a useful way to guide discussion at the consensus meetings. Scoring was impacted by the reviewers' interpretation of qualitative methodologies and relevance of articles. They also suggested that this more open approach allowed articles that contained important information or sections to be reviewed in spite of low scores. Some papers contained very brief methods sections, which made the scoring difficult and resulted in low scores.

Acknowledging that synthesizing diverse data in reviews is a challenge, Goldsmith, Bankhead and Austoker (2007) developed a new review method. They implemented it on a review of evidence to inform guidelines for the content of patient information related to cervical screening. They followed standard procedures for conducting a systematic review, including:

- a priori study design
- comprehensive literature search
- two people independently reviewing titles and abstracts
- quality assessment
- data extraction

The researchers used checklists to perform quality assessment on both quantitative and qualitative studies. Established checklists were used for quantitative studies (Critical Appraisals Skills Program (CASP), Scottish Intercollegiate Guidelines Network (SIGN), New Zealand Guidelines Group (NZGG) and UK Government Chief Social Researcher's Office (UKGCSRO)) (CASP, 2005; Khan et al., 2001; Lethaby, Wells, & Furness, 2001; Spencer, Ritchie, & Lewis, 2004) . The researchers developed a checklist for the qualitative studies that included purpose, population, response rate, outcome definition and assessment. All checklists included a comments section. The methodological quality of studies was rated as ++ (all or most criteria were fulfilled), + (some criteria were fulfilled) or – (few or no criteria were fulfilled). The quality appraisal provided insight into the strengths and weaknesses of each study. However, methodological flaws did not result in studies being excluded. Quality scores were incorporated in the data synthesis phase of the review. Goldsmith et al. (2005) incorporated the synthesis guidelines set out in the GRADING system (Atkins et al., 2005) discussed in the Critical Appraisal Tools section of this paper.

DeWalt et al. (2004) conducted a systematic review of studies that measured literacy plus one or more health outcomes. The eligible study designs were observational: prospective and retrospective cohort studies, case-control studies and cross-sectional studies. Their review followed a standard systematic review process, including:

- a comprehensive literature search of several electronic databases;
- well-defined research questions;
- inclusion/exclusion criteria;
- two authors reviewing full articles
- quality assessment and reporting of findings.

The researchers adapted the quality assessment tool from West et al. (2002). Each study was rated according to:

- study population
- comparability of subjects
- validity and reliability of the literacy measure
- maintenance of comparable groups
- outcome measure
- statistical analysis
- controlling for confounding

Based on these criterions, the studies received a ‘good,’ ‘fair’ or ‘poor’ rating. This tool has not been validated. Although the review authors were able to provide a statistical analysis of the literacy and health outcomes, they cautioned that the primary study data posed many challenges due to the various reading measures used and cut points for analysis. As well, lack of adequate statistical measures, inadequate controlling for confounders and lack of adjustment for multiple comparisons made comparisons between studies difficult.

Thomson et al. (2006) examined data reporting on socioeconomic determinants of health in the UK to determine if governmental investment in the area had improved health. The data from 10 evaluations were synthesised using systematic review methods, including:

- a search strategy
- a priori inclusion/exclusion criteria
- two people independently reviewing articles
- data extraction
- data synthesis

Eight of the impact evaluations used case studies where data were gathered from a few sites to represent the national program. Methodological issues with the evaluation reports included poor evaluation methods and data sources and low sample sizes. The authors used a narrative synthesis to report findings. They reported that this approach was a new way to build evidence with the message tailored to impact the development of health public policy. The synthesis was challenged by the methodological shortcomings of the included studies.

Two studies (Stein, Dalziel, Garside et al., 2005; Dalziel, Round, Stein et al., 2005) examined the quality of evidence of effectiveness used in health technology assessments (HTA). Case series, commonly used in HTAs, constitute a weak form of evidence in the hierarchy of evidence. In ideal situations, case series data would not be used in systematic reviews; however, there are times when they might be the only available evidence. Finding no simi-

lar studies in their literature search, Stein et al. examined a sample of systematic reviews of case series that explored the association between the characteristics of case series and outcome. They did not compare case series results with RCT results. In their analysis, they found little evidence of association between methodological features, sample size or prospective approach and outcome. They provided a narrative report of their findings, and included a table of individual outcomes of the analysed variables. They cautioned that all the interventions studied were surgical, which might limit the generalizability of their findings. In a second study, Dalziel et al. compared results from RCTs and case series studies in surgical interventions. They examined reports that included data from both RCTs and case series. They compared these studies using the intervention arm of the RCTs as a comparator: meta-analysis of RCT was compared with weighted robust regressions using the intervention as the confounding factor and estimating the coefficient size. Their analysis revealed that the RCTs showed no outcome difference between treatment types, while the case series showed an increase in mortality of 1–2% between treatments. Limitations of this review include:

- analysis was constrained by methodological flaws in the case series studies, such as poor reporting of data;
- use of specific surgical interventions may limit the generalizability of findings;
- findings were based on a small number of studies.

In a report that examined evidence for, and methods of, evaluating non-randomized trials, Deeks et al. (2003) identified 194 tools used to assess the quality of these trials. Sixty tools covered at least five of six pre-specified domains for internal validity and were classified as ‘top tools.’ Fourteen tools were ranked as ‘best tools,’ covering three of four core items of particular importance for non-randomized studies (allocation, comparable groups, prognostic factors identified and use of case-mix adjustment). The authors identified six of 14 tools as being suitable for use in systematic reviews. The strength of those tools was the phrasing of items that channelled the reviewer’s responses in a systematic way to ensure the assessments were as objective as possible.

Katrak et al. (2004) produced a systematic review of 121 published critical assessment tools from 108 papers located by searching electronic databases and the Internet. Most of the tools (87%) were specific to research design, with many (45%) of those developed for experimental studies (RCTs and CCTs). Of a total of 16 generic critical appraisal tools, six were developed for experimental and observational studies. Eleven tools were found to be useful for any qualitative and quantitative research design. The authors extracted 74 items from 19 critical appraisal tools for observational studies. These items focused on:

- data analyses
- consideration of confounders
- sample size or power calculation
- whether appropriate statistical analysis was undertaken

They extracted 36 items from the seven critical appraisal tools for qualitative studies. Most of the items focused on assessing external validity, methods of data analyses and justification of the study. These tools did not contain items about sample selection, randomization,

blinding, intervention or bias. The generic tools, reportedly usable for either experimental or observational studies, contained items that focused on sampling selection and data analyses, such as appropriateness of statistical analyses and sample size and power calculation. This review found no gold standard appraisal tool for any type of study.

Section Summary

These systematic reviews successfully incorporated primary studies with data that did not have control groups, or they examined studies with critical appraisal tools for the quality assessment of studies without control groups. The reviews that examined the feasibility of combining the results of RCTs and non-RCTs suggest that well-designed studies can produce similar results, regardless of the study type. Many of these reviews suggest that poor execution of the primary study design, rather than the study design itself, result in lower quality scoring. They site under-reporting of methodology, poor management of confounders and/or lack of consistent analytical terminology as problematic when trying to compare data between studies.

Methods for Synthesizing Quantitative Studies without Control Groups

This section includes five papers that described methods for synthesizing quantitative studies without control groups.

In a 2005 paper, Greenhalgh et al. described their process of meta-narrative review, developed as a methodology for the synthesis of evidence across disciplines. They used a six-step process that included planning, searching for literature, mapping, doing an appraisal, synthesizing and making recommendations. According to the authors, their approach of producing “storied” accounts of the key research traditions moved complex literature out of confusion and into “sensemaking.” Five key principles underpinned this meta-narrative technique:

- pragmatism
- pluralism
- historicity
- contestation
- peer review

This approach may be helpful in situations where the scope of the research is broad and the literature diverse, and where researchers have approached a common problem using different study designs.

Whittemore and Knafl (2005) highlighted the integrative review method as a good option for integrating evidence from studies that emerge from diverse methodologies. The authors outlined strategies to enhance the rigour of integrative reviews. They proposed five steps to help ensure the methodological rigour of this type of review:

1. Problem identification: Ensures the research question is clearly defined.
2. Literature search: Incorporates a comprehensive search strategy.

3. Data evaluation: Becomes somewhat complex due to the methodologically diverse primary studies included in the review. Evaluating quality cannot follow the standard systematic review process, but may need to focus on examples of authenticity, methodological quality, informational value and representativeness of available primary studies.
4. Data analysis: Includes data reduction, display, comparison and conclusions. Data reduction is achieved through an overall classification system in which primary studies are divided into subgroups that provide a logical order for analysis. This subgroup classification can be based on types of evidence, chronology, setting, sample characteristics or by a predetermined conceptual classification. The next step in the data reduction process is to code and extract data into a manageable framework. Data display involves converting the extracted data by specific variables and subgroups. These displays include matrices, graphs and charts that allow for comparison across studies. Data comparison identifies patterns, themes and relationships between and amongst the variables. The final stage in data analysis is drawing conclusions.
5. Presentation: Includes the reporting of findings in an integrative review that can be in the form of tables or diagrams, and should include implications for practice, policy and research.

Norris and Atkins (2005) examined 49 reviews released during a five-year period by the Evidence-based Practice Centers that included evidence from studies using designs other than RCTs. Those designs included cohort studies and time series, before-after case series studies and cross-sectional studies. The authors suggested that one benefit to incorporating cohort studies and case series studies in reviews is that these types of designs are better able to capture long-term outcomes than are RCTs. They observed that while most of the 49 reviews incorporated non-randomized studies, the evidence tended to be from the RCTs. Assessing the quality of non-randomized trials was a challenge. Of the 49 reviews, 25% did not assess quality, 16% used published checklists or scoring systems, 10% adapted published scoring instruments and 49% used checklists that had been developed by the reviewers. Few of the tools had been tested for validity. To ensure adherence to principles of best evidence and to reduce bias, Norris and Atkins suggested that reviews begin with a detailed protocol which would then be strictly followed from review inception to completion. As well, the search strategy may need to be more fluid and repeated to compensate for the lack of sensitive and specific search strategies used for RCTs. The authors made the following recommendations:

1. Reviewers should assess the availability of RCTs addressing their review question before determining final inclusion criteria.
2. When considering the inclusion of non-randomized trials, reviewers need to consider potential biases and whether those biases can be minimized in well-conducted non-randomized trials.
3. The review process must be transparent so that reasons for inclusion and exclusion of study designs are explicit.
4. Reviewers should consider quality assessment of individual studies.

5. Reviewers should be aware of the impact of their inclusion criteria on the findings, and include a discussion of the potential impacts of bias on their conclusions.
6. Methodological quality of the included studies needs to be part of the discussion and conclusions, regardless of study design.

Jackson and Waters (2005) looked specifically at the challenges faced in writing systematic reviews for the public health sector. Some of the inherent difficulties included multi-component interventions, multiple outcomes measured, diverse populations and mixed study designs. They agreed with others who say that in public health, the intervention and the population need to dictate the study design, not vice versa. The researchers acknowledged the controversy around the appropriateness of the systematic review process for public health intervention studies. They also offered suggestions for improving the process as a means of moderating that criticism. Searching for literature is complex and time consuming. Therefore, it is essential to allow sufficient time for an adequate search, as well as to search multiple electronic databases and sources of grey literature. Quality assessment of all included articles is necessary, and they recommended using the tool developed by the Effective Public Health Practice Project, Hamilton, Ontario (Thomas et al., 2004a) "Quality Assessment Tool for Quantitative Studies." This strong instrument (Deeks et al., 2003) can be used on RCTs, quasi-experimental and uncontrolled studies. Attention to theoretical frameworks in both primary studies and systematic reviews can help explain the differences that occur between the planning and the outcomes of the interventions. Measuring the integrity of inventions can help determine if an intervention was ineffective because of poor planning, or because the delivery of the intervention was incomplete. Due to the diverse populations receiving public health interventions, researchers should expect heterogeneity. Researchers and reviewers will not find all these components readily incorporated in all primary studies, but awareness of these components may help to strengthen systematic reviews emerging from public health intervention research.

Atkins and DiGuiseppi's 1998 paper examined research needs about preventative health services use. Evidence from RCTs was not always available in the area of health promotion and preventative health. The researchers cited the known problems associated with observational studies as reason for caution; however, they acknowledged that these studies can add to the body of knowledge. The strengths of observational studies include:

- establishing linkages in the causal pathway;
- building understanding of the natural history of disease;
- identifying risk factors;
- measuring compliance with and adverse effects of treatments;
- determining accuracy of diagnostic tests;
- assessing efficacy of interventions.

Risk of bias can be modified with representative, population-based samples and careful consideration of potential confounders. Researchers also need to recognise that observational studies tend to exaggerate beneficial effects of treatment. The authors recommend including a large number of well-designed studies with consistent and preferably large effects when using observational studies to build treatment recommendations.

Section Summary

Synthesizing and integrating quantitative data from studies without control groups is challenging. These six papers reported on various methods to incorporate such data in a useful format. The authors agreed that there is benefit to incorporating evidence from diverse study designs, but quality assessment is paramount for providing trustworthy evidence. Primary studies that clearly identify potential biases and probable impacts on effect can be usefully incorporated into systematic reviews.

Combining Qualitative and Quantitative Studies

This section examines three articles that describe methods for synthesizing data from both qualitative and quantitative studies.

Determining the quality of non-experimental studies is challenging and controversial; however, these types of studies often contain important information for public health. Assessing quality using standard tools is not particularly helpful because few of these studies would score in the high or moderate category. Harden et al. (2004) used a standard approach for methodological quality assessment in systematic reviews on 35 primary non-intervention studies. Only four met all seven criteria used to assess quality. They developed a data extraction tool to help deconstruct each study (both quantitative and qualitative). From this, they reconstructed the results and presented them in structured summaries and evidence tables. The synthesis process was non-linear and involved two reviewers going back and forth between the papers, the data extraction and the evidence tables. Pooling results, as done in meta-analysis, was not appropriate. Instead, these researchers used aggregate methods according to identified themes to synthesize findings. To answer the main research question, the reviewers integrated the findings from the non-intervention studies with 36 intervention studies, comparing the results to identify similarities and differences (Oliver et al., 2005). Using these two types of results allowed for greater breadth of perspectives and deeper understanding of public health issues from the point of view of the people who receive the targeted interventions.

Meta-reviews, as described by Wong and Raabe (1996), incorporate both qualitative and quantitative data. The authors suggested using a meta-review to address some of the limitations associated with traditional qualitative reviews, such as subjectivity in article selection and weights assigned to individual studies, and lack of quantitative analysis. This was achieved by including the meta-analysis of quantitative data within a narrative review. The first factor for determining whether a study should be included is the study design; studies with different designs should not be grouped together for meta-analysis. Meta-analysis can increase statistical power, but it cannot compensate for other study design methodological issues. There are some basics steps in conducting a meta-review:

1. Define the research question.
2. Conduct a comprehensive literature review.
3. Create a priori inclusion criteria.
4. Conduct a traditional qualitative review.

5. Conduct a quantitative meta-analysis.
6. Integrate a traditional qualitative review with quantitative meta-analysis.
7. Apply criteria for causation in interpretation.

The authors suggest that incorporating a meta-analysis with a traditional narrative review can reduce subjectivity.

A 2007 paper by Sandelowski, Barroso, & Voils, discussed the need for health disciplines to include both qualitative and quantitative research findings in reviews, especially in light of the growth of mixed methods research design. The authors examined 42 reports, including journal articles, unpublished dissertations/theses and technical reports. To differentiate their approach from a standard systematic review, no report was excluded based on methodological quality. To synthesize their findings, they developed a qualitative meta-synthesis model that grouped primary research findings in topical and thematic units. The meta-summary included the extracting, grouping and formatting of findings, as well as the calculating of frequency and effect sizes for the quantitative data. This approach can be used to synthesize mixed methods surveys in which the data collection and analysis procedures are similar.

Section Summary

Incorporating qualitative and quantitative data in one report can be helpful for building an understanding of the success or failure of interventions. However, assessing the quality of studies without control groups can be problematic, especially if quality is assessed with traditional methods used in systematic reviews or meta-analysis. Results of quality assessment were not used as exclusion criteria, but were instead incorporated in the discussion of findings. Grouping findings according to a thematic and aggregate method is an appropriate alternative approach to meta-analysis. Researchers could incorporate some systematic review processes, such as having two people review and perform data extraction, but most authors suggest using a narrative approach for presenting results when meta-analysis is not appropriate.

Studies Comparing Results Using Differing Methods

The search included two studies that examined issues that arise when comparing results of randomized and non-randomized studies of the same intervention.

Jefferson and Demicheli (1999) assessed the capability of different research designs for testing four aspects of vaccine performance (immunogenicity, duration of immunity conferred, incidence and seriousness of side effects, and number of infections prevented by vaccination). Their findings indicated that in vaccinology, experimental and non-experimental study designs are frequently complementary. However, in some situations, vaccine quality could only be measured with one type of study. For instance, an RCT is the preferred study design to measure side effects. Non-experimental designs are appropriately applied when:

- an experiment is impossible, unnecessary or inappropriate;
- the individual efficacy is to be measured in terms of infrequent adverse events;
- when interventions prevent rare events or the population effectiveness of an inter-

vention to be measured in the long term.

Using interview data collected from 17 authors or users of Health Technology Assessments (HTA), Rotstein and Laupacis (2004) shed light on the differences between HTAs and systematic reviews. These differences include:

1. Methodological standards—HTAs may include literature of poor methodological quality if a topic is important to decision-makers.
2. Replication of previous studies—systematic reviews do not need to be repeated if previous studies were of high-quality or when there is no new high-quality evidence; in HTAs there is often a need to repeat studies to defend the report's conclusions.
3. Choice of topics—topics are more policy-oriented with HTAs, while systematic reviews tend to be driven by effectiveness questions.
4. Inclusion of content experts (in systematic reviews) and policy-makers (in HTAs) as authors.
5. Inclusion of economic evaluations—in HTAs.
6. Making policy recommendations—in HTAs.
7. Dissemination of report—more often actively done for HTAs.

HTAs are not specifically designed for evaluating scientific evidence, while systematic reviews are designed for this purpose. Yet the impartial nature of scientific investigation would help strengthen the policy decisions emerging from HTA evidence. Different levels of evidence (below RCTs in the hierarchy) are more acceptable in HTAs.

Section Summary

These two comparisons of results from different designs about the same intervention point out that it is useful to consider different designs in reviews, not only because they may be the best available information, but also because they can more likely answer questions related to long-term effectiveness, adverse events and rare events. The strengths of including other designs outweigh the limitations.

Discussion

The main research focus for this paper was the methods of including results from non-randomized trials, as well as the quality and synthesis of that data. This is an important research question because in many areas of health and public health, evidence from randomized control trials is not available. Historically, evidence emerging from non-randomized trials has been criticized as unreliable due to a greater potential of bias. It is commonly understood that there is a greater tendency for selection, allocation or attrition bias in non-randomized trials. Readers of data from non-randomized trials need to be aware that these biases may exist, and researchers should account for these potential biases in their studies. Some suggestions emerging from the studies examined in this paper include:

- There is a need for a strong and transparent study design in all non-randomized trials.
- The research question should determine the study design.
- Study authors need to incorporate detailed information about the study; lack of information means that readers cannot determine what has or has not been done in the study design.
- There is a need for consistency in terminology and descriptions across non-randomized studies to facilitate comparisons.

There are several methodological approaches for including data from non-randomized studies. We found the work of Whittemore and Knafl (2005) to be most systematic. Whittemore and Knafl provide a five-step integrative review method that incorporates problem identification, literature search, data evaluation, data analysis and presentation of data.

Many researchers use various assessment tools to determine the quality of the included studies. Norris and Atkins (2005) provide a comprehensive list of recommendations for assessing study quality that emerged from the 49 reviews they examined. These recommendations were outlined earlier in this paper. Other authors used existing checklists or scoring instruments, such as MINORS or GRADING. Some adapted tools to conform to the study designs found in the primary studies, while others developed new tools that reflected the information to be extracted from included studies. Many of these tools have not been tested for validity or reliability. We recommend that further research be done to determine the validity and reliability of these quality assessment tools for non-randomized trials. Where possible, reviewers should use tools that have been tested and determined to be valid and reliable for non-randomized trials, such as the Quality Assessment Instrument for Primary Studies developed by the Effective Public Health Practice Project, Hamilton, Ontario.

Including data from non-randomized trials is challenging; however, several benefits can be achieved from its inclusion. Drawing from non-randomized studies means that researchers can add a wider body of literature in their search strategies (Thomas et al., 2004). Incorporating these studies may broaden the scope of the questions that can be answered. This research can shed light on why, how and for whom interventions or programs succeed or fail (Oliver et al., 2005).

Recommendations

There is currently no standardized model for synthesizing the results of studies that do not have control groups. This paper examined synthesis methods, critical appraisal tools and studies that deal with appraising and synthesizing the results from quantitative studies that lack control groups.

This paper outlined some precautions to take when using non-controlled studies as evidence, and made recommendations for processes and tools:

1. Incorporate studies without control groups into systematic reviews, especially when there are no other studies to consider. Studies without control groups can also provide information on long-term effectiveness, rare events and adverse effects.
2. Do not generalize the direction of differences in outcome between randomized and non-controlled studies. While effect sizes are more often smaller in randomized controlled trials (RCTs), some show the reverse or the same results across study designs.
3. Use Whittemore and Knafl's (2005) approach to include results of non-randomized trials. Whittemore and Knafl provide a five-step integrative review method that incorporates problem identification, literature search, data evaluation, data analysis and presentation of data.
4. Recommendations for methods of systematic review including studies without control groups:
    a) Use a critical appraisal tool that has been tested for internal consistency, test-retest and inter-rater reliability, and at least face and criterion validity (Downs & Black, 1998; Slim, Nini, Forestier et al., 2003; Steuten, Vrijhoef, van-Merode et al., 2004; Thomas, Ciliska, Dobbins et al., 2004a; Zaza, Wright-De Aguero, Briss et al., 2000).
    b) Do not meta-analyse results from observational studies.
5. Suggestions to authors of primary studies:
    a) Ensure a strong and transparent study design in all non-randomized trials.
    b) Confirm that the research question determines the study design.
    c) Incorporate detailed information about the study; lack of information means that readers cannot determine what has or has not been done in the study design.
    d) Ensure consistency in terminology and descriptions across non-randomized studies to facilitate comparisons.
6. Implications for further research:
    a) Conduct reliability and validity testing of any critical appraisal tools.
    b) Conduct further studies to assess the different results obtained by randomized versus non-randomized trials.
7. Including data from non-randomized trials is challenging; however, several benefits can be achieved:

a) A wider body of literature can be included in search strategies (Thomas, Harden, Oakley et al., 2004).

b) The scope of the questions that can be answered may be broadened.

c) This research can shed light on why, how and for whom interventions or programs succeed or fail (Oliver, Harden, Rees et al., 2005).

Study	Purpose	Methods/Results	Comments/Issues
Rangel, et al. 2003 USA ***CRITICAL APPRAISAL TOOLS***	To describe the development and potential applications of a standardized quality assessment scale designed for retrospective studies in pediatric surgery.	Developed a comprehensive quality assessment instrument incorporating 30 items within three subscales. The subscales were designed to assess clinical relevance, reporting methodology and the strength of conclusions. Global quality ratings (poor, fair or good) were derived by combining scores from each subscale. Six independent reviewers examined inter-rater reliability by assessing the instrument in 10 retrospective studies from pediatric surgery literature. Findings: • The instrument had excellent inter-rater reliability. • There were high levels of agreement for all items within instrument (84.6% concordance, n = 1573 items) and for all individual subscales (range, 73.3% to 85.8%, n = 60 to 1258)	Conclusions: • The authors have developed a standardized and reliable quality assessment scale for the analysis of retrospective data in pediatric surgery. • Results may not be generalizable to other interventions. Potential applications: • Provide practicing surgeons with a knowledge base to critically evaluate published retrospective data; • Provide standardized methodology for SR of existing retro data; • Develop standardized reporting guidelines for use in peer-reviewed journals.
Margetts, et al. 1995 UK ***CRITICAL APPRAISAL TOOLS***	To develop a scoring system to help judge the scientific quality of observational epidemiologic studies linking diet with risk of cancer.	The scoring system was developed from key headings used in developing research protocols and included questions under headings: • three for case control studies (dietary assessment, recruitment of subjects and analysis) – scoring system in Appendix A; • four for cohort studies (dietary assessment, definition of cohort, ascertainment and analysis) – scoring system in Appendix B. Inter-observer variation was assessed: There was good agreement between observers in the ranking of studies. The scoring system was applied to a review of meat and cancer risk: From what seemed like a large literature sample, relatively few studies scored well (defined as a score > 65%). However, these studies tended to provide more consistent information. For case control studies, 34 out of 106 studies scored > 65%. For cohort studies, 10 out of 41 studies scored > 65%.	The prototype scoring system reported here helped the authors describe the epidemiologic data on diet and cancer, but it is not generalizable to other settings or interventions.

Study	Purpose	Methods/Results	Comments/Issues
Downs & Black 1998 UK ***CRITICAL APPRAISAL TOOLS***	To test the feasibility of creating a valid and reliable checklist with the following features: Appropriate for assessing both randomized and non-randomized studies. Provides both an overall score for study quality and a profile of scores, not only for the quality of reporting, internal validity and power, but also for external validity.	The study began with a pilot version of the tool. The researchers asked three experts to evaluate face and content validity. Afterward, two raters used the tool to assess 10 randomized and 10 non-randomized studies to measure reliability. Findings: • Using different raters, the checklist was revised and tested for internal consistency (Kuder-Richardson 20), test-retest and inter-rater reliability, criterion validity and respondent burden. • The researchers found high internal consistency in Quality Index. • Test-retest and inter-rater reliability of the Quality Index were good.	Conclusions: • It is possible to produce a checklist that evaluates randomized and non-randomized studies. • The main concern is with the checklist's external validity and application of it to topics beyond the study's purpose.
Slim, et al. 2003 France ***CRITICAL APPRAISAL TOOLS***	To develop and validate a Methodological Index for Non-Randomized Studies (MINORS) which could be used by readers, manuscript reviewers or journal editors to assess the quality of such studies.	The index consisted of 12 items. In the case of non-comparative studies, items included: • a stated aim of the study; • inclusion of consecutive patients; • prospective collection of data; • endpoint appropriate to the study aim; • unbiased evaluation of endpoints; • follow-up period appropriate to the major endpoint; • loss to follow-up not exceeding 5%. In the case of comparative studies, items included: • a control group having the gold standard intervention; • contemporary groups; • baseline equivalence of groups; • prospective calculation of the sample size; • statistical analyses adapted to the study design. The index had good inter-reviewer agreement, high test-retest reliability by the kappa-coefficient and good internal consistency by Cronbach's alpha coefficient.	MINORS is a valid instrument designed to assess the methodological quality of non-randomized surgical studies, whether comparative or non-comparative. Researchers may need to use caution if applying the instrument to non-surgical interventions.

Study	Purpose	Methods/Results	Comments/Issues
Zaza et al. 2000 US ***CRITICAL APPRAISAL TOOLS***	1. To assist with consistency, reduce bias and improve validity for researchers using the Guide to Community Prevention Services: Systematic Reviews and Evidence-Based Recommendations (the Guide) 2. To develop a standardized data extraction form to use with the Guide.	The form was developed by: • reviewing methodologies from other systematic reviews; • reporting standards established by health and social science journals; • examining evaluation, statistical and meta-analysis literature; • soliciting expert opinions. The form was used to assess the methodological quality of primary studies (23 questions) and to classify and describe key characteristics of the intervention and evaluation (26 questions). The researchers examined study procedures and results and assessed threats to validity across six categories: • intervention and study descriptions • sampling • measurement • analysis • interpretation of results • other execution issues	This process improved the validity and reliability of the Guide. The standardized data extraction form can assist researchers and readers of primary studies to review the content and quality of the studies. The form can also assist in the development of manuscripts for submission to peer reviewed journals.
Greer, et al. 2000 US ***CRITICAL APPRAISAL TOOLS*** ***(evidence grading systems)***	To describe in detail the evidence and conclusion grading system developed by the Institute for Clinical Systems Improvement (ICSI) for use by the practicing clinicians who write the documents and use them in making decisions about patient care	The grading system included an evaluation of individual research reports and an assessment of the overall strength of the evidence supporting a particular conclusion or recommendation. The rating system to assess the quality of individual research reports can be quickly understood and more easily used by practicing physicians. The intent of the grading system is to highlight a report that is unusually strong or markedly flawed. The evidence grading system: • includes the conclusion grading worksheet, which calls for statement of a conclusion, a summary of research reports that support or dispute the conclusion, assignment of classes and quality markers to the research reports, and assignment of a grade to the conclusion; • has been used in the writing of more than 40 guidelines and numerous technology assessment reports.	Conclusions: • The system appears to be successful in overcoming the complexity of some published systems of grading evidence, while still yielding a defensible classification of conclusions based on the strength of the underlying evidence. • The system's reliability needs to be rigorously tested.

Study	Purpose	Methods/Results	Comments/Issues
Steuten, et al. 2004 Netherlands ***CRITICAL APPRAISAL TOOLS***	To describe to what extent existing instruments are useful in assessing the methodological quality of Health Technology Assessment (HTA) of disease management.	Problems were identified in the assessment of the methodological quality of six HTAs of disease management with three different instruments. The problems mainly concerned: • study design (RCT versus observational studies); • criteria for selection and restriction of patients; • baseline and outcome measures (several parameters are necessary); • blinding of patients and providers (which is impossible in disease management); • description of complex, multifaceted interventions and avoidance of co-interventions. The researchers proposed and validated a new instrument, HTA-DM, that: • includes four components (study population, description of intervention, measurement of outcomes and data-analysis/presentation of data). • contains 15 items (on internal validity, external validity, statistical considerations).	
Chou & Helfand 2005 USA ***METHODS*** ***(Systematic Reviews** (SR) **of harms)*** ***CRITICAL APPRAISAL TOOLS***	To review the methodological challenges in performing SRs of harms and highlight examples of approaches to them from 96 EPC evidence reports.	Some studies have found that observational studies and RCTs report similar estimates of effects. Others have found that clinical trials report higher risks for adverse events than observational studies (possibly due to poorer assessment of harm in observational studies, or they may be more likely to be published if they report good methods/ results). Including large databases may provide useful information about harm. Data from practice-based networks are often richer in clinical detail than administrative databases—it may be possible to identify and measure likely confounders with more confidence, but they are harder to find using electronic searches because many such analyses are proprietary. Including case reports can help identify uncommon, unexpected or long-term adverse drug events that are often different from those detected in clinical trials. Assessing the quality of harm reporting: • Several SRs found that prospective or retrospective design, case-control compared with cohort studies, and smaller compared with larger case series had no clear effect on estimates of harm.	Better data about harm is needed to conduct balanced SRs. Further research is needed to empirically determine the impact of including data from different sources on assessments of harm, and further develop and test criteria for rating the quality of harm reporting. It is important to: • avoid inappropriate statistical combinations of data; • carefully describe the characteristics and quality of included studies; • thoroughly explore potential sources of heterogeneity.

Study	Purpose	Methods/Results	Comments/Issues
McIntosh, et al. 2004 UK ***METHODS*** ***(SRs)***	To present the experience of conducting systematic reviews (SRs) of harmful effects and to make suggestions for future practice and further research.	The authors evaluated the methods used in three SRs, focusing on the review question, study designs and quality assessment. Review question: • One review had a specific question, focused on providing information on specific harmful effects to furnish an economic model; the other two reviews had broader questions. Study designs: • All three included randomized and observational data, but each defined the inclusion criteria differently. Quality assessment: • Applied published checklists—encountered problems; • Inadequate reporting of basic design features of the primary studies—checklists omit key features, such as how harmful effects data were recorded.	Key areas for improvement include: • focusing the review question, as broad and non-specific questions are not helpful to the process of review; • developing standardized methods for the quality assessment of studies of harmful effects.
Atkins, et al. 2005 USA ***CRITICAL APPRAISAL TOOLS*** ***(grading system)***	To pilot test and further develop the GRADE approach to grading evidence and recommendations.	The GRADE approach was developed by the GRADE Working Group to grade the quality of evidence and the strength of recommendations. Researchers used 12 evidence profiles in this pilot study. Each profile was made based on information available in a systematic review. 17 people independently graded the level of evidence and the strength of recommendations for each of the 12 evidence profiles. Quality of evidence for each outcome: • Researchers found that in addition to study design, quality, consistency and directness, other quality criteria also influenced judgments about evidence; sparse data, strong associations, publication bias, dose response and situations where all plausible confounders strengthened rather than weakened confidence in the direction of the effect. Reliability: • There was a varied amount of agreement on the quality of evidence for the outcomes relating to each of the twelve questions (kappa coefficients for agreement beyond chance ranged from 0 to 0.82). • There was fair agreement about the relative important of each outcome. • There was poor agreement about the balance of benefits and harms and recommendations (could, in part, be explained by the accumulation of all the previous differences in grading of the quality and importance of the evidence). • Most of the discrepancies were easily resolved through discussion.	

Study	Purpose	Methods/Results	Comments/Issues
Khan, et al. 2001 UK ***CRITICAL APPRAISAL TOOLS***	To provide background information on study quality and describe how to develop quality assessment checklists.	The researchers described reasons to include study quality assessment in reviews, types of biases, some methods to protect against bias in primary effectiveness studies and how quality assessment instruments should be developed. Quality criteria for assessment of observational studies includes answering a series of questions for cohort studies, case-control studies and case series. Cohort studies: • Is there sufficient description of the groups and the distribution of prognostic factors? • Are the groups assembled at a similar point in their disease progression? • Is the intervention/treatment reliably ascertained? • Were the groups comparable on all important confounding factors? • Was there adequate adjustment for the effects of these confounding variables? • Was a dose-response relationship between intervention and outcome demonstrated? • Was outcome assessment blind to exposure status? • Was follow-up long enough for the outcomes to occur? • What proportion of the cohort was followed-up? • Were drop-out rates and reasons for drop-out similar across intervention and unexposed groups? Case-control studies: • Is the case definition explicit? • Has the disease state of the cases been reliably assessed and validated? • Were the controls randomly selected from the source of population of the cases? • How comparable are the cases and controls with respect to potential confounding factors? • Were interventions and other exposures assessed in the same way for cases and controls?	Conclusions: • In a review that incorporates these study designs, it can be difficult to determine if differences are a result of the intervention or group incompatibilities. • Results should be viewed with caution.

Study	Purpose	Methods/Results	Comments/Issues
Thomas, et al. 2004 Canada ***CRITICAL APPRAISAL TOOLS*** ***METHODS*** ***(SRs)***	To describe four issues related to systematic literature reviews of the effectiveness of public health nursing interventions: 1. process of systematically reviewing the literature; 2. development of a quality assessment instrument; 3. the methods/ results of the EPHPP to date; 4. some methods/ results of the dissemination used.	Domains for assessing observational studies: • study question • study population • comparability of subjects • exposure/intervention • outcome measure blinding • statistical analysis • results • discussion • funding Components of the EPHPP Quality Assessment Tool: • selection bias • design • confounders • blinding • data collection methods • withdrawals and drop-outs	This model of scoring has been tested as a reliable and valid tool. This scoring system indicates a preference for RCTs and CCTs, which will tend to score higher than non-randomized studies. The best score that can be achieved by observational studies is "moderate," and only if all other components of the study are very well designed.
Ramsay, et al. 2003 USA ***CRITICAL APPRAISAL TOOLS*** ***METHODS*** ***(interrupted time series** (ITS) **designs)***	To critically review the methodological quality of ITS (interrupted time series) designs using studies included in two systematic reviews (a review of mass medic interventions and a review of guideline dissemination and implementation strategies).	Quality criteria for ITS designs: • Intervention occurred independently of other changes over time. • Intervention was unlikely to affect data collection. • The primary outcome was assessed blindly or was measured objectively. • The primary outcome was reliable or was measured objectively. • The composition of the data set at each time point covered at least 80% of the total number of participants in the study. • The shape of the intervention effect was pre-specified. • A rationale for the number and spacing of data points was described. • The study was analysed appropriately using time series techniques. Findings: • A total of 66% (n = 38) of ITS studies did not rule out the threat that another event could have occurred at the point of intervention. • Reporting of factors related to data collection, the primary outcome and completeness of dataset were generally done in both reviews. • All the studies were considered "effective" in the original report, but approximately half of the re-analysed studies showed no statistically significant differences.	Interrupted time series designs are often analysed inappropriately, underpowered and poorly reported in implementation research.

Study	Purpose	Methods/Results	Comments/Issues
Conn & Rantz 2003 USA ***CRITICAL APPRAISAL TOOLS*** ***METHODS*** ***(quality and study outcomes)***	1. To discuss ways that researchers conceive of and assess quality. 2. To determine associations between study quality and outcomes. 3. To determine strategies for managing the varied quality of primary studies in a meta-analysis.	Assessing methodological quality: • No single standard exists for addressing the quality variations in primary studies. • Table 1 – summary of commonly noted components of intervention research quality. • Instruments to measure quality – more than 100 primary study quality scales exist for measuring the quality of primary studies, but they vary dramatically in size, composition, complexity and extent of development. • Quality measurement scales have developmental and application problems. • Most scales result in a single, overall score for quality. Only a few scales contain subscales that profile strengths and weaknesses. • Instruments have proven difficult to apply consistently, even in RCTs. Relationship between quality and study outcomes: • Findings have often been contradictory. • The findings are mixed on whether low-quality studies under- or over-estimate effect sizes compared to high quality studies. • Different scales generate diverse assessments of study quality, which cause inconsistency in efforts to relate study quality to outcome. Strategies to manage quality: • Set minimum levels for inclusion or require that certain quality attributes be present—can involve setting inclusion criteria, selecting a priori particular quality-scale cut-off scores; limitation: excluding research that may be of lower rigour goes against the scientific habit of examining data, letting the data speak. • Weight effect sizes by quality scores— allows inclusion of diverse studies, but relies on questionable quality measures. • Consider quality to be an empirical question—can examine associations between quality and effect sizes and thus preserve the purpose of meta-analysis to systematically examine data; limitation: problems with scale measures of overall quality may limit confidence in findings related to overall quality measures.	Conclusions: • Researchers are increasingly combining strategies to overcome the limitations of using a single approach. • Further work to develop valid measures of primary study quality dimensions will improve the ability of meta-analyses to inform research and nursing practice.

Study	Purpose	Methods/Results	Comments/Issues
Linde, et al. 2002 Germany ***STUDIES*** ***(randomized vs. non-randomized)***	To investigate: 1. Possible differences in patients, interventions, design-independent quality aspects, and response rates of randomized vs. non-randomized studies of acupuncture for chronic headache. 2. If non-randomized studies provide relevant additional info (i.e., methods/ results on long-term outcomes, prognostic factors, adverse effects, response rates). 3. Possible explanations if response rates in randomized and non-randomized patients differ.	Quality criteria: • Is the sampling method clear and clearly described? • Is there a clear headache diagnosis? • Are patients characterized (at least age, sex, duration, severity of symptoms)? • Is there at least a four-week baseline period? • Are there at least two clinical headache outcomes? • Is a headache diary used? • Are co-interventions described? • Are at least 90% of patients included analysed after treatment? • Are at least 80% of patients treated analysed at early (<6 months) follow-up? • Are at least 80% of patients treated analysed at late (≥6 months) follow-up? Methods/Results: • 59 studies were included: 24 randomized trials, and 35 non-randomized studies (five non-randomized controlled cohort studies, 10 prospective uncontrolled studies, 10 case series and 10 cross sectional surveys). • 10 of the 24 RCTs and 26 of the 35 non-randomized studies met less than five quality criteria. • A total of 535 patients received acupuncture treatment in the 24 RCTs compared to 2695 in the 35 non-randomized studies. • On average, randomized trials had smaller sample sizes, met more quality criteria and had lower response rates (0.59; 0.48-0.69) vs. 0.78 (0.72-0.83). • Randomized or not, studies meeting more quality criteria had lower response rates. • Non-randomized studies didn't have significantly longer follow-up periods. Of the three that included an analysis of prognostic variables, only one reported on adverse effects, and the degree of generalizability was unclear.	

Study	Purpose	Methods/Results	Comments/Issues
MacLehose, et al. 2000 UK ***STUDIES*** ***(RCTs vs. Quasi-experimental and observation*** *(QEO))*	To compare estimates of effectiveness from RCTs vs. QEOs and examine the association between methodological quality and magnitude of effectiveness estimates.	***Strategy 1: Compare RCT and QEO study estimates of effectiveness of any intervention, where both estimates were reported in a single paper.*** 14 papers with 38 comparisons were reviewed; 25 were of low quality and 13 were of high quality. Discrepancies between RCT and QEO study estimates of effect size and outcome frequency for intervention and control groups were smaller for high- than low-quality comparisons. Conclusion: • QEO study estimates of effectiveness may be valid if important confounding factors are controlled for. • Low quality comparisons tend to have more extreme QEO study estimates of effect. • Methods/results are limited by the number of papers reviewed (few) and potentially unrepresentative nature of evidence reviewed. ***Strategy 2: Compare RCT and QEO study estimates of effectiveness for specified interventions, where the estimates were reported in different papers.*** Interventions: mammographic screening (MSBC) of women to reduce mortality from breast cancer; folic acid supplementation (FAS) to prevent neural tube defects in women trying to conceive. 34 papers were reviewed (17 on MSBC, 17 on FAS). Eight and four papers, respectively, were individually or cluster assigned RCTs, five and six were non-randomized trials or cohort studies, and three and six were matched or unmatched case–control studies. Two studies, one of MSBC and one of FAS, used some other study design. Both cohort and case-control studies had lower total quality scores than RCTs; cohort studies also had significantly lower scores than case-control studies. Meta-regression of study attributes against relative risk estimates showed no association between effect size and study quality for either intervention. Estimates from RCTs and cohort studies were not significantly different, but case-control studies gave significantly different estimates for both MSBC (greater benefit) and FAS (less benefit).	Recommendations: • Standards for reporting quasi-experimental and observational studies should be developed. • Innovative search strategies should be explored. • Methods for identifying studies that provide a direct comparison of estimates from randomized and non-randomized data are needed.

Study	Purpose	Methods/Results	Comments/Issues
Britton 1998 UK ***STUDIES*** ***(RCTs vs. non-RCTs)***	To explore the issues related to the process of randomization that may affect the validity of conclusions drawn from the Methods/Results of RCTs and non-randomized studies	Previous comparisons of RCTs and non-randomized studies: • 18 papers that directly compared the Methods/Results of RCTs and prospective non-randomized studies were found and analyzed • Neither the RCTs nor the nonrandomized studies consistently gave larger or smaller estimates of the treatment effect. • The type of intervention did not appear to be influential Exclusions: • The number of eligible subjects included in the RCTs ranged from 1% to 100%. • Reasons for exclusions may be medical (e.g. high risk of adverse events in certain groups) or scientific (selecting only • small homogeneous groups in order to increase the precision of estimated treatment effects); also blanket exclusions are common in RCTs Participation: • Most RCTs failed to document adequately the characteristics of eligible individuals who did not participate in trials. Adjusting for baseline differences: • in non-randomized studies: adjustment for differences often changed the treatment effect size but not significantly; importantly, the direction of change was inconsistent.	Conclusions: • Methods/Results of RCTs and non-randomized studies do not inevitably differ • The available evidence suffers from many limitations but it suggested that it may be possible to minimize any differences by ensuring that subjects included in each type of study are comparable.
Lemmer, et al. 1999 UK ***METHODS*** ***(SRs)***	To report on key problems and difficulties encountered in a systematic literature review that, in the absence of RCTs, drew upon theoretical studies of decision-making and practice-based research.	A total of 164 reviews of 82 publications were carried out. A modified version of the Cochrane Collaboration Protocol was used: • Databases and keywords used for search strategies were selected in consultation with a librarian specializing in health and related subjects. • Omitted sections were considered inappropriate for reports that were not RCTs, and were replaced with headings for a broader range of methodologies. There was also a section at the end of the protocol for reviewers' comments on the appropriateness and significance of each review. A numerical score was given to each article; 8 is the maximum score and 1 is the minimum score. This scale was used to indicate the methodological rigour and relevance of each article.	Qualitative studies provide little information on methodology, making quality assessment difficult. • All reviewers agreed, however, that some of the articles that were rated poor contained important issues or sections whose omission would be a detriment to the study. Reaching a consensus score was difficult as individual reviewers interpreted articles and research designs in different ways.

Study	Purpose	Methods/Results	Comments/Issues
Goldsmith, et al. 2007 UK ***METHODS***	To describe the review methods developed and the difficulties encountered during the process of updating a systematic review of evidence to inform guidelines for the content of patient information related to cervical screening.	Studies of women in an organized, systematic cervical screening program were included; all study designs, except for opinion pieces, were included. 233 studies were retrieved for further evaluation. Data extraction was conducted for 79 studies (52 quantitative and 27 qualitative). 32 studies were reported as relevant. 13 qualitative studies, seven non-comparative descriptive studies, three RCTs, three quasi-RCTs, three cross sectional studies and one each of cluster RCT, retrospective case control study, and non-comparative time series study were included in the report. The quality of each individual study was assessed using established checklists: SIGN, CASP, NZGG, UKGCSRO and a reviewer checklist. The basic principles of the GRADE approach were applied to the synthesis of the quantitative evidence.	Conclusions: • The review questions were best answered by evidence from a range of data sources. • Qualitative research was often highly relevant and specific to many components of the screening information materials.
DeWalt, et al. 2004 USA ***METHODS***	To review the relationship between literacy and health outcomes. Review of Studies – examines observational studies that reported original data measuring literacy with any valid instrument and measured one or more health outcomes.	Included studies with outcomes related to health and health service and measurement of literacy skills with a valid instrument (West et al., 2002). Graded each study according to: • adequacy of study population • comparability of subjects • validity and reliability of the literacy measurement • maintenance of comparable groups • appropriateness of the outcome measurement • appropriateness of statistical analysis • adequacy of control of confounding	The average quality of the included studies was fair to good. The authors found that most studies failed to adequately address confounding and the use of multiple comparisons.
Thomson, et al. 2006 UK ***METHODS***	To synthesize data on the key socioeconomic determinants of health and health inequalities reported in evaluations of the national UK regeneration program.	Included evaluations that reported achievements drawing on data from at least two target areas of a national urban regeneration program, or area-based initiatives (ABIs) in the UK. Nineteen evaluations reported impacts on health or socioeconomic determinants of health. Data was synthesized from 10 evaluations. Standard systematic review procedures were used, including: • comprehensive search strategy; • a priori inclusion/exclusion criteria; • two people independently reviewing articles; • data extraction; • data synthesis.	Methodological shortcomings of the primary studies challenged the data synthesis process. The authors suggested that interventions and programs attached to a theoretical framework (e.g., theory of change) will help to build evidence.

Study	Purpose	Methods/Results	Comments/Issues
Stein, et al. 2005 USA ***METHODS*** ***(health technology assessments*** *(HTAs))*	To investigate the association between frequency and methodological characteristics in a sample of health technology assessments.	The researchers included reviews on four interventions in three reports. Data was extracted from published and unpublished reports and was extracted by one reviewer and checked by a second. Various regression analysis tests were undertaken on the data. The study outcome measures were reported at different time periods, depending on the length of follow-up of the entire study. The review looked at: • sample size • prospective/retrospective approach • multi- or single-centre organizations • consecutive recruitment • independence of outcome measure • length of follow-up • publication date Findings: • Little evidence was found of an association between methodological characteristics and outcome. • Sample size and prospective approach were not shown to be associated with outcome.	All outcomes were surgical interventions, which may impact generalizability. The small number of cases and limited number of studies in each set of case series may limit the precision and generalizability. Case series may be particularly prone to publication bias.

Study	Purpose	Methods/Results	Comments/Issues
Dalziel, et al. 2005 UK ***METHODS*** ***(case series in HTAs)*** ***STUDIES*** ***(case series vs. RCTs)***	1. To review the use of case series (CS) in National Institute for Clinical Excellence (NICE) HTA reports. 2. To systematically review the methodological literature for papers relating to the validity of aspects of case series design. 3. To investigate characteristics and findings of case series using examples from the UK's HTA program	Review of use of case series in NICE HTAs: Of 47 completed HTAs, 14 included information from CSs. Inclusion criteria for CSs included study size and length of follow-up. There was no consensus on which CSs to include in HTAs, how to use them or how to assess their quality. Systematic review of methodological literature: A search was conducted to find studies that assessed aspects of case series design, analysis or quality in relation to study validity. No empirical studies were found. Investigation of characteristics and findings of case series designs: No relationship was found between sample size and outcome frequency or between prospective data collection and outcome frequency. One analysis each (in different topic areas) found a significant association between multi-centre studies and outcome, between independent outcome measurement and outcome frequency and between earlier publication and outcome frequency. Length of follow-up was found to be significantly associated with outcome frequency in three analyses. Comparison between case series and RCTs: Compared with RCT evidence, which showed no difference between PTCA and CABG, case series estimates of mortality showed a 1–2% increase in mortality for CABG. For angina recurrence, neither case series nor RCT data showed any difference between the two interventions.	The included studies were all surgical interventions which might limit generalizability to other interventions or settings. The study drew on a small number of cases, therefore results should be viewed with caution.

Study	Purpose	Methods/Results	Comments/Issues
Deeks, et al. 2003 UK ***CRITICAL APPRAISAL TOOLS*** ***(evaluate ability of case-mix adjustment to control for bias)***	To consider methods and related evidence for evaluating bias in non-randomized intervention studies.	Eight studies compared the results of randomized and non-randomized studies across multiple interventions. A total of 194 tools were identified for assessing non-randomized studies. Findings: • The results from non-randomized studies sometimes differ from the results of randomized studies of the same intervention. • Many quality assessment tools exist, but appraisal of studies omits key quality domains. • Non-randomized studies should only be undertaken when RCTs are not feasible.	
Katrak, et al. 2004 Australia ***CRITICAL APPRAISAL TOOLS***	To summarize content, intent, construction and psychometric properties of published, currently available critical appraisal tools to identify common elements and their relevance to allied health research.	A total of 121 published critical appraisal tools were included in the SR, sourced from 108 papers. Findings: • 87% of tools were specific to research design, with most tools having been developed for experimental studies (38% of all tools sourced). • There was great variability in items contained in the critical appraisal tools. • 12% (n = 14 instruments) of available tools were developed using specified empirical research. • 49% of the tools summarized quality appraisal into a numeric summary score. • Few tools had documented evidence of validity of their items, or reliability of use. • Guidelines regarding administration of tools were provided in 43% (n = 52) of cases.	Conclusions: • There is no “gold standard” critical appraisal tool for any study design, nor is there any widely accepted generic tool that can be applied equally-well across study types. • Consumers of research should carefully select tools. • Selected tools should have published evidence of empirical basis for their construction, validity of items and reliability of interpretation, and guidelines for use.

Study	Purpose	Methods/Results	Comments/Issues
Greenhalgh, et al. 2005 UK ***METHODS*** ***(meta-narrative review)***	To introduce/describe and report on the development of a new method: meta-narrative review.	Meta-narrative review was developed as the methodological base for the synthesis of evidence across multiple disciplinary fields. It has particular strengths as a synthesis method when: • the scope of a project is broad and the literature diverse; • different groups of scientists have asked different questions and used different research designs to address a common problem; • 'quality' papers have different defining features in different literature; • there is no universally agreed-upon process for pulling the different bodies of literature together. The study involved a systematic mapping phase to collect and compare the different overarching storylines of the rise and fall of diffusion research that is judged relevant to the overall research question. Phases of meta-narrative review: • planning • searching • mapping • appraisal • synthesis • recommendations Five key principles underpin the meta-narrative technique: pragmatism, pluralism, historicity, contestation and peer review.	The method should be tested prospectively. Its contribution to the mixed economy of methods for the systematic review of complex evidence should be explored further.
Whittemore & Knafl 2005 USA ***METHODS*** ***(integrative review)***	To distinguish the integrative review method from other review methods and to propose methodological strategies specific to the integrative review method to enhance rigour.	Issues described and strategies used to enhance the rigour of the integrative review method in nursing: • Well-specified review purpose and variables of interest—help to accurately operationalize variables and extract appropriate data from primary sources. • Well-defined literature search strategies—identify the maximum number of eligible primary sources, using at least two to three strategies. • Data evaluation stage—since integrative review method includes diverse primary sources, the complexity of evaluating the quality increases; how quality is evaluated will vary depending on the sampling frame; in a review with a diverse sampling frame inclusive of empirical and theoretical sources, an approach similar to historical research to evaluate quality may be appropriate. • Data reduction—divide the primary sources into subgroups according to some logical system to facilitate analysis. • Data display, data comparison, conclusion drawing and verification, presentation.	The integrative review method allows for the combination of diverse methodologies (e.g., experimental and non-experimental research) and has the potential to play a greater role in evidence-based practice.

Study	Purpose	Methods/Results	Comments/Issues
Norris & Atkins 2005 USA ***METHODS*** ***(SRs)***	To examine the use of non-randomized studies in evidence-based practice center (EPC) reports, addressing questions of the effectiveness of treatment interventions.	Of the 107 EPC reports released between Feb 1999 and Sept 2004, 78 examined at least one question of efficacy or effectiveness of a clinical intervention. 49 of the reports included evidence from study designs other than RCTs; these reports examined pharmacotherapy, medical devises, surgery, complementary and alternative and behavioural interventions. Challenges in using non-randomized studies in systematic reviews: Terminology to describe different non- randomized study designs is inconsistent in clinical research literature and EPC reports. There are no established guidelines as to when non-randomized studies can or should be considered for inclusion in systematic reviews, or what study designs to consider. It is difficult to assess the quality of non-randomized studies. Of 49 reports that included non-randomized study designs: • 12 (25%) didn't assess study quality; • 16% used a checklist or scoring system; • 10% adapted a previously published instrument; • the remainder (49%) used instruments that the reviewers had developed themselves.	Recommendations for systematic reviewers: • Assess the availability of RCTs before determining final inclusion criteria. • Consider the pros and cons of non-RCT designs. • Provide a rationale for decisions to include/exclude specific study designs. • Assess important domains of study quality. • Discuss how including various study designs may affect conclusions. • Discuss how the quality of studies and the body of evidence may affect conclusions. Recommendations for researchers: • Minimize sources of bias regardless of study design. • Examine the effects of various sources of bias on measured outcomes. • Use consistent terminology for study design. • Devise comprehensive search strategies for non-randomized study designs.
Jackson & Waters 2005 Australia ***METHODS*** ***(SRs)***	To provide recommendations to reviewers on the issues to address within a public health systematic review and to indirectly provide advice to researchers on the reporting requirements of primary studies for the production of high quality systematic reviews.	To ensure reviews meet the needs of users, establish an advisory group (members should be familiar with topic area, useful to include perspectives of policy makers, funders, practitioners, potential users). Issues to address when reviewing public health interventions: • inclusion of study designs • search for public health literature • quality assessment • theoretical frameworks for intervention • integrity of interventions • heterogeneity • integration of qualitative and quantitative studies • ethics and inequalities – health interventions effective in reducing inequalities and improving health of marginalized/ disadvantaged	Recommendations: • Use the Quality Assessment Tool for Quantitative Studies developed by Thomas et al., 2004.

Study	Purpose	Methods/Results	Comments/Issues
Atkins & Di-Guiseppi 1998 USA ***METHODS*** ***(research on preventive health care)***	Drawing on experiences from the US Preventive Services Task Force, to outline some major areas where research is needed to define the appropriate use of specific preventive services (e.g., screening tests, counselling interventions, immunizations, chemoprophylaxis).	Observational studies: • help to establish linkages in causal pathway; • help to understand natural history of disease and identify risk factors, measuring compliance with and adverse effects of treatments; • help to determine accuracy of diagnostic tests; • help to assess efficacy of interventions; • have an inherent bias.	Recommendations based solely on observational evidence require high-quality studies showing consistent and preferably large effects. Where only a few observational studies of adequate quality are available or effect sizes are modest or inconsistent, additional studies in different populations would be a valuable addition to the literature.
Harden, et al. 2004 UK ***METHODS*** ***(views studies)***	To describe the methods developed for reviewing research on people's perspectives and experiences ("views" studies) alongside trials, within a series of reviews on young people's mental health, physical activity and healthy eating.	Two types of studies were reviewed: 1. Intervention studies – to identify effective, ineffective and harmful interventions. 2. Non-intervention studies – to describe factors associated with mental health, physical activity and healthy eating. "Views" studies: • 35 met the inclusion criteria. The studies varied in the methods used—many could not easily be classified as "qualitative" or "quantitative." Most failed to meet seven basic methodological reporting standards used in a newly developed quality assessment tool. The benefits of bringing together view studies in a systematic way include: • gaining a greater breadth of perspectives and a deeper understanding of public health issues from the point of view of those targeted by interventions; • helping reflect on study methods that may distort, misrepresent or fail to pick up people's views; • creating greater opportunities for people's own perspectives and experiences to inform policies to promote their health.	Emerging framework reported in subsequent publication (Oliver, et al., 2005)

Study	Purpose	Methods/Results	Comments/Issues
Wong & Raabe 1996 USA ***METHODS*** ***(meta-review)***	To provide some basic guidelines for non-epidemiologists to evaluate meta-analyses in occupational cohort studies.	Some problems and limitations of traditional qualitative reviews: • Selection of studies was usually not subject to pre-determined criteria. • Weights assigned to individual studies were usually subjective. • No quantitative summaries of risk estimates were done. Findings: • Rather than replacing qualitative reviews, quantitative meta-analysis should be made part of the overall assessment. • The term "meta-review" is proposed to emphasize the importance of both qualitative and quantitative components in a comprehensive review process, meaning it includes a meta-analysis. Basic steps in conducting a meta-review: 1. Define the research question. 2. Conduct a literature search. 3. Determine criteria for inclusion. 4. Conduct a traditional qualitative review. 5. Conduct a quantitative meta-analysis. 6. Integrate traditional qualitative review with quantitative meta-analysis. 7. Apply criteria for causation in interpretation.	Benefits of meta-review: • useful in selecting studies, in organizing, presenting and summarizing methods/ results from individual studies; • can also be used to detect heterogeneity among studies. Major benefits of conducting a meta-analysis include: • enhancing statistical power (especially when original studies are small); • providing a properly weighted summary risk estimate; • taking consistency of study and methods/results into consideration; • minimizing the problem of multiple comparisons; • examining data for heterogeneity.
Sandelowski, et al. 2007 USA ***METHODS*** ***(meta-summary)***	To address the challenge of managing the differences presumed to exist between qualitative and quantitative research, and advance the use of qualitative meta-summary as a technique useful for synthesizing qualitative and quantitative descriptive findings.	42 reports (35 journal articles, six unpublished theses or dissertations and one technical report) were included. Of the 42 reports, 12 were qualitative studies, three were intervention studies, one was a mixed methods study and 26 were varieties of quantitative observational studies. Features of qualitative meta-summaries include: • a quantitatively-oriented aggregation approach to research synthesis; • the extraction, grouping and formatting of findings, and the calculation of frequency and intensity effect sizes; • able to synthesize methods/results of qualitative and quantitative surveys of responses obtained from similar data collection and analysis procedures; • able to conduct a posteriori analyses of the relationship between reports and findings.	In contrast with most systematic review methods, this study is biased toward inclusion rather than exclusion of studies.

Study	Purpose	Methods/Results	Comments/Issues
Jefferson & Demichelli 1999 UK ***STUDIES*** ***(experimental vs. non-experimental)***	To examine the relation between experimental and non-experimental study designs in vaccinology.	The authors described experimental and non-experimental studies and their approaches. They assessed each study design's capability of testing four aspects of vaccine performance, namely immunogenicity, duration of immunity conferred, incidence and seriousness of side effects and number of infections prevented by vaccination.	Non-experimental designs should be applied when: • an experiment is impossible because the aim of evaluation is to assess vaccine effectiveness; • an experiment is unnecessary (e.g., Smallpox vaccination evaluation); • an experiment is inappropriate (trial population not large enough to detect event/outcome, reluctance/refusal to participate, ethical/legal/political obstacles); • individual efficacy is measured in terms of an infrequent adverse event; • interventions prevent rare events; • the population effectiveness of the vaccine is to be measured in terms of long-term, rare and serious consequences of disease.
Rotstein & Laupacis 2004 Canada ***STUDIES*** ***(HTAs vs. SRs)***	To elucidate the important differences between health technology assessments (HTAs) and systematic reviews (SR).	The researchers interviewed 17 authors or users of HTA. They identified seven areas of differences between HTAs and SRs: 1. methodological standards (HTAs may include literature of relatively poor methodological quality if a topic is of importance to decision-makers) 2. replication of previous studies (relatively common for HTAs but not systematic reviews) 3. choice of topics (more policy-oriented for HTAs, while systematic reviews tend to be driven by researcher interest) 4. inclusion of content experts and policy-makers as authors (policy-makers more likely to be included in HTAs, although there are potential conflicts of interest) 5. inclusion of economic evaluations (more often with HTAs, although economic evaluations based upon poor clinical data may not be useful) 6. making policy recommendations (more likely with HTAs, although this must be done with caution) 7. dissemination of the report (more often actively done for HTAs)	HTAs are not solely concerned with the evaluation of scientific evidence.

Reference List

Scottish Intercollegiate Guidelines Network 2001–2005. (2001). *SIGN 50: A guideline developers' handbook*. Retrieved February 1, 2009 from http://www.sign.ac.uk/guidelines/fulltext/50/index.html.

Critical Appraisal Skills Programme (CASP) and evidence-based practice. (2005). *Critical Appraisal Skills Programme: Making sense of evidence - CASP appraisal tools*. Oxford, Public Health Resource Unit, Milton Keynes Primary Care NHS Trust. Retrieved February 1, 2009 from http://www.phru.nhs.uk/Pages/PHD/resources.htm.

Atkins, D., Briss, P. A., Eccles, M., Flottorp, S., Guyatt, G. H., Harbour, R. T., et al. (2005). Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system. *BMC Health Services Research, 5*, 25.

Atkins, D., & DiGuiseppi, C. G. (1998). Broadening the evidence base for evidence-based guidelines. A research agenda based on the work of the U.S. Preventive Services Task Force. American *Journal of Preventive Medicine, 14*, 335-344.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine, 342*, 1878-1886.

Britton, A., McKee, M., Black, N., McPherson, K., Sanderson, C., & Bain, C. (1998). Choosing between randomised and non-randomised studies: A systematic review. *Health Technology Assessment, 2*, 1-124.

Chou, R., & Helfand, M. (2005). Challenges in systematic reviews that assess treatment harms. *Annals of Internal Medicine, 142*, 1090-1099.

Conn, V. S., & Rantz, M. J. (2003). Focus on research methods. Research methods: Managing primary study quality in meta-analysis. *Research in Nursing & Health, 26*, 322-333.

Dalziel, K., Round, A., Stein, K., Garside, R., Castelnuovo, E., & Payne, L. (2005). Do the findings of case series studies vary significantly according to methodological characteristics? *Health Technology Assessment, 9*, iii-iv, 1.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*, 1-173.

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., & Pignone, M. P. (2004). Literacy and health outcomes: A systematic review of the literature. *Journal of General Internal Medicine, 19*, 1228-1239.

DiCenso, A., Prevost, S., Benefield, L., Bingle, J., Ciliska, D., Driever, M., et al. (2004). Evidence-based nursing: Rationale and resources. *Worldviews on Evidence-Based Nursing, 1,* 69-75.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*, 377-384.

Goldsmith, M. R., Bankhead, C. R., & Austoker, J. (2007). Synthesising quantitative and qualitative research in evidence-based patient information. *Journal of Epidemiology and Community Health, 61*, 262-270.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science & Medicine, 61*, 417-430.

Greer, N., Mosser, G., Logan, G., & Halaas, G. W. (2000). A practical approach to evidence grading. *Joint Commission Journal on Quality Improvement, 26*, 700-712.

Harden, A., Garcia, J., Oliver, S., Rees, R., Shepherd, J., Brunton, G., et al. (2004). Applying systematic review methods to studies of people's views: An example from public health research. *Journal of Epidemiology and Community Health, 58*, 794-800.

Jackson, N., & Waters, E. (2005). Criteria for the systematic review of health promotion and public health interventions. *Health Promotion International, 20*, 367-374.

Jefferson, T., & Demicheli, V. (1999). Relation between experimental and non-experimental study designs. HB vaccines: A case study. *Journal of Epidemiology and Community Health, 53,* 51-54.

Katrak, P., Bialocerkowski, A. E., Massy-Westropp, N., Kumar, S., & Grimmer, K. A. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4:22*.

Khan, K., ter Riet, G., Popay, J., Nixon, J., & Kleijnen, J. (2001). Stage II: Conducting the review. Phase 5: Study quality assessment. In: K. S. Khan, G. ter Riet, J. Glanville, et al. *Undertaking systematic reviews of research on effectiveness. CRD's guidance for carrying out or commissioning reviews 2nd Edition* (pp. 1-20). York: NHS Centre for Reviews and Dissemination, University of York.

Lemmer, B., Grellier, R., & Steven, J. (1999). Systematic review of non-random and qualitative research literature: Exploring and uncovering an evidence base for health visiting and decision making. *Qualitative Health Research, 9,* 315-328.

Lethaby, A., Wells, S., & Furness, S. (2001). *Handbook for the preparation of explicit evidence-based clinical practice guidelines*. Aukland, New Zealand: New Zealand Guidelines Group, Effective Practice Institute of the University of Auckland.

Linde, K., Scholz, M., Melchart, D., & Willich, S. N. (2002). Should systematic reviews include non-randomized and uncontrolled studies? The case of acupuncture for chronic headache. *Journal of Clinical Epidemiology, 55*, 77-85.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment, 4*, 1-154.

Margetts, B. M., Thompson, R. L., Key, T., Duffy, S., Nelson, M., Bingham, S., et al. (1995). Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. *Nutrition and Cancer, 24*, 231-239.

McIntosh, H. M., Woolacott, N. F., & Bagnall, A. M. (2004). Assessing harmful effects in systematic reviews. *BMC Medical Research Methodology, 4*, 19.

Norris, S., & Atkins, D. (2005). Challenges in using nonrandomized studies in systematic reviews of treatment interventions. *Annals of Internal Medicine, 142*, 1112-1119.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. (2005). Systematic reviews of health effects of social interventions: 2. Best available evidence: How low should you go? *Journal of Epidemiology and Community Health, 59*, 886-892.

Oliver, S., Harden, A., Rees, R., Shepherd, J., Brunton, G., Garcia, J., et al. (2005). An emerging framework for including different types of evidence in systematic reviews for public policy. *Evaluation, 11*, 428-446.

Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003). Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care, 19,* 613-623.

Rangel, S. J., Kelsey, J., Colby, C. E., Anderson, J., & Moss, R. L. (2003). Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery, 38*, 390-396.

Rotstein, D., & Laupacis, A. (2004). Differences between systematic reviews and health technology assessments: A trade-off between the ideals of scientific rigor and the realities of policy making. *International Journal of Technology Assessment in Health Care, 20*, 177-183.

Sandelowski, M., Barroso, J., & Voils, C. I. (2007). Using qualitative metasummary to synthesize qualitative and quantitative descriptive findings. *Research in Nursing & Health, 30*, 99-111.

Slim, K., Nini, E., Forestier, D., Kwiatkowski, F., Panis, Y., & Chipponi, J. (2003). Methodological index for non-randomized studies (minors): Development and validation of a new instrument. *ANZ Journal of Surgery, 73*, 712-716.

Spencer, L., Ritchie, J., & Lewis, J. (2004). *Quality in qualitative evidence: A framework for assessing research evidence.* (2nd ed.) London: Government Chief Social Researcher's Office.

Stein, K., Dalziel, K., Garside, R., Castelnuovo, E., & Round, A. (2005). Association between methodological characteristics and outcome in health technology assessments which included case series. *International Journal of Technology Assessment in Health Care, 21*, 277-287.

Steuten, L. M., Vrijhoef, H. J., van-Merode, G. G., Severens, J. L., & Spreeuwenberg, C. (2004). The health technology assessment – Disease management instrument reliably measured methodologic quality of health technology assessments of disease management. *Journal of Clinical Epidemiology, 57*, 881-888.

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004a). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*, 176-184.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., et al. (2004b). Integrating qualitative research with trials in systematic reviews. *British Medical Journal, 328*, 1010-1012.

Thomson, H., Atkinson, R., Petticrew, M., & Kearns, A. (2006). Do urban regeneration programmes improve public health and reduce health inequalities? A synthesis of the evidence from UK policy and practice (1980–2004). *Journal of Epidemiology and Community Health, 60*, 108-115.

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). Systems to rate the strength of scientific evidence. *Evidence Report – Technology Assessment (Summary)*, 1-11.

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*, 546-553.

Wong, O., & Raabe, G. K. (1996). Application of meta-analysis in reviewing occupational cohort studies. *Occupational and Environmental Medicine, 53*, 793-800.

Zaza, S., Wright-De Aguero, L. K., Briss, P. A., Truman, B. I., Hopkins, D. P., Hennessy, M. H., et al. (2000). Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *American Journal of Preventive Medicine, 18*, 44-74.