
STAGE II

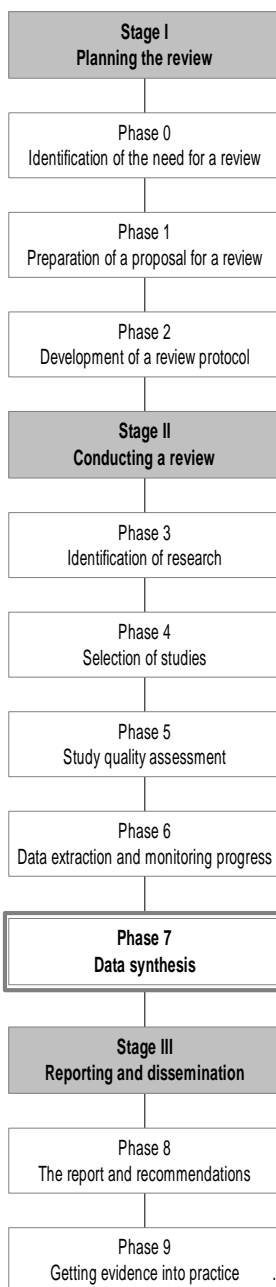
Conducting the review

PHASE 7

Data synthesis

**Jon Deeks, Khalid S Khan, Fujian Song,
Jennie Popay, John Nixon & Jos Kleijnen**

This phase describes synthesis of extracted data. There are two components to synthesis. Non-quantitative synthesis involves tabulation of study characteristics and results to summarise their findings. This approach allows a qualitative assessment of the evidence. This is supplemented by quantitative synthesis (use of statistical methods for assessing variation in results and generating pooled results) if appropriate.



2.7.1	Aims of data synthesis	2
2.7.2	Descriptive or non-quantitative synthesis	2
2.7.3	Quantitative synthesis	4
2.7.4	Data synthesis of effectiveness studies	5
2.7.5	Data synthesis of test accuracy studies	14
2.7.6	Publication bias	16
2.7.7	Data synthesis of qualitative research	18
2.7.8	Data synthesis of economic evaluations	21
2.7.9	Key points	24
2.7.10	References	25

2.7.1 Aims of data synthesis

The aim of data synthesis in a systematic review is to collate and summarise the results of included primary studies.¹⁻³ It can be achieved through a descriptive or non-quantitative synthesis complemented by the use of formal statistical techniques (such as meta-analysis) if appropriate. In addition to generating a summary of the effects of interventions, it is an integral part of data synthesis to investigate whether the effects are consistent across the included studies, and if not, to investigate reasons for the apparent differences. This phase includes core information about issues in data synthesis (see Box 7.1): full details of the suggested methods may be found in the referenced materials.

2.7.2 Descriptive or non-quantitative synthesis

The objective of a descriptive or non-quantitative review is to collate and present the extracted data in a manner such that information about the characteristics (population,

Box 7.1

Key concepts in data synthesis for systematic reviews

Descriptive data synthesis

A non-quantitative synthesis of the collated evidence to assess the extent of the evidence and to plan the quantitative synthesis. It allows a qualitative assessment of variation in study characteristics, quality and results (heterogeneity). In some situations where there are numerous studies with consistent and large effects, it may be possible to discern effects solely from this synthesis.

Quantitative data synthesis

A synthesis using a group of statistical techniques to combine the results of the included studies (meta-analysis), to assess heterogeneity, and to quantitatively evaluate other aspects like publication bias. Meta-analysis is used to calculate a pooled estimate of effect and its confidence interval.

Heterogeneity

The variability or differences between studies in terms of key characteristics (clinical heterogeneity), quality (methodological heterogeneity) and effects (heterogeneity of results). Statistical tests of heterogeneity may be used to assess whether the observed variability in study results (effect sizes) is greater than that expected to occur by chance.

Homogeneity

The degree to which the studies included in a review are similar. Studies are considered statistically homogeneous if their results vary no more than might be expected by the play of chance.

Sensitivity analysis

An analysis used to determine how the results of a systematic review change due to variations arising from uncertain decisions or assumptions about the data and the methods that were used.

Publication bias

A bias in the research literature where the likelihood of publication of a study is influenced by the significance of its results. For example, studies in which an intervention is not found to be effective may be less likely to be published. Systematic reviews that fail to identify such studies may overestimate the true effect of an intervention. In some subject areas (e.g. in alternative medicine), studies showing effectiveness may also suffer from publication bias.

interventions, outcomes and study quality) and results of the studies included in the review, are summarised in a meaningful way. This is best done by tabulation, which allows readers to look at the evidence, its methodological rigour and the differences between studies. The descriptive overview is an essential part of data synthesis on which an understanding of the data, planning the quantitative data synthesis and preventing errors in its interpretation are based (see Box 7.2).

The process of carrying out the descriptive part of data synthesis should be explicit and rigorous.^{4, 5} Decisions about how data are to be grouped and tabulated should be based on the questions that the review is addressing (see Phase 1.2.2). The effectiveness of a health care intervention is likely to depend on a large number of factors (known and unknown) relating to who receives it, who delivers it and how, and in what context. Studies evaluating the impact of the intervention may differ in relation to these factors. The key elements in the descriptive approach to data synthesis may include the following characteristics:

- a) population
- b) interventions
- c) settings where the technology was applied
- d) environmental, social and cultural factors that may influence compliance
- e) nature of the outcomes measures used, their relative importance and robustness
- f) the validity of the evidence
- g) the sample sizes and the results of the studies included in the review.

These factors should be summarised succinctly in the tables. The tables should be structured to highlight the similarities and the differences between included studies. From a critical analysis of these tables, it should be possible to assess qualitatively if there are differences between studies in key characteristics of the participants, interventions or outcome measures (clinical heterogeneity); the study designs and quality (methodological heterogeneity); and the reported effects (heterogeneity in results). Thus, it should be possible to decide whether the studies are similar enough

Box 7.2

Descriptive data synthesis in systematic reviews

- Describes the studies
- Allows an assessment of whether participants, interventions and outcomes in the studies allow the generalisation of the results of the review, or whether there are restrictions and omissions which restrict the applicability of the results
- Allows an assessment of whether the quality of the studies is adequate to trust their results
- Demonstrates the absence of data for some planned comparisons
- Demonstrates the degree of clinical heterogeneity in patients and interventions prohibiting planned comparisons
- Demonstrates heterogeneity of outcomes preventing use of the data for meta-analysis
- If a formal quantitative meta-analysis cannot be done, allows an assessment of whether the effect of the treatment is large enough to be regarded as obvious, and if the effects are consistent across the studies.

for it to be sensible to calculate an average estimate of effectiveness. In some cases important factors or variables may not have been reported in the included studies. The non-quantitative synthesis should also highlight the problems arising from the lack of important information.

Data synthesis involves computation of an average effect where the results of each study are weighted according to some measure of the study's importance. Each study's weight usually relates to its sample size and the resulting precision of the estimate of effect. Statistical methods of meta-analysis are explicit numerical formulations of this process, and should be used wherever possible, as described below. However, in some circumstances the required numerical values may not be available. When the treatment effects are large and consistent it may still be possible to estimate a qualitative effect indicating whether the treatment does good or harm and the range of possible effect. Such a qualitative synthesis should also take account of the statistical significance of the study results, the study quality, and the relative importance of the difference studies.

Where there are important differences between the studies in terms of participants, interventions, outcomes and methods that are thought potentially to relate to study results, it is usually not sensible to estimate an overall average effect. However, in some situations subgroups of similar studies can be identified from the tabulations for which an average effect could be computed, or variables identified which could be explored as potential explanations of statistical heterogeneity. Thus the descriptive part of the synthesis can help plan investigations of heterogeneity. Although it is more credible to follow analyses outlined in the protocol, it is necessary to ensure that adequate data are available for such analyses, and to be aware whether there are additional issues of importance, which were not known of when the protocol was produced.

2.7.3 Quantitative synthesis

An assessment of the tabular summaries helps in planning the quantitative syntheses by highlighting the comparisons that could be made, the outcomes that could be combined (meta-analysis) and the study characteristics that should be considered when investigating variation in effects (heterogeneity). It is within the framework of this descriptive overview that any quantitative synthesis should take place.

First, it should be determined whether quantitative synthesis is at all possible and if so whether it would be appropriate. Meta-analysis is not possible when the necessary data to perform a meta-analysis cannot be obtained and it may not be appropriate when the data are sparse or when the studies are too heterogeneous to be sensibly combined. Heterogeneity may arise when the populations in the various studies have different characteristics, when the delivery of the interventions is variable, or when studies of different designs and quality are included in the review. Where substantial heterogeneity is present, a non-quantitative synthesis may informally explore how the differences in study characteristics affect their results.

Once it is established that a meta-analysis is possible and appropriate, reviewers have to make three choices before beginning. First, which comparisons should be made? Second, which outcome measures should be used in the synthesis? Third, which effect measure (a measure of association quantifying the effect of intervention) should be used to describe effectiveness? These issues should have been considered in the protocol. The nature of

the comparisons and the outcome measures should be related directly to the questions being posed in the review and the main comparisons should have been specified (see Phase 1.2.7). However, it might become necessary to modify or delete certain comparisons and outcomes, and/or add new ones in light of the descriptive review. For example, it may only become apparent after data collection that identified studies do not contribute enough data for the particular comparisons and outcome measures outlined in the protocol. Also the extent to which study results are similar enough to be sensibly analysed together can only be discovered after the data have been collected. Therefore, the ultimate choice of comparisons and outcome measures will be influenced by the findings of the non-quantitative synthesis.

2.7.4 Data synthesis of effectiveness studies

This section focuses on the general principles and methods of synthesis of studies of effectiveness. The quantitative methods described here are for synthesis of studies with randomised controlled design but the same principles apply to comparative observational studies.

2.7.4.1 Choice of effect measure

There are four issues of importance in selecting an effect measure. What type of data is the outcome measure? Is the measure interpretable by those who will use the review? Is the measure likely to be consistent across the studies and transferable? Does the measure have the mathematical properties required to give a valid answer?

There are three types of data commonly encountered in systematic reviews. Dichotomous or binary data are where each individual must be in one of two states, such as dead or alive. Such data can be summarised using odds ratios, risk ratios or risk differences. Continuous data are outcomes that are summarised as means, arising through measurements or the use of assessment scales, and are summarised in systematic reviews as differences in means, or standardised differences in means (effect sizes). Survival or time to event data is commonly encountered in cancer therapies and some other fields, where the outcome of principal interest is the time to the occurrence of an event. It is usually summarised using hazard ratios. Some outcome measures do not fit this classification. For example, some outcomes may be short ordinal scales, such as pain scales (where individuals' rate their pain as none, mild moderate or severe), for which it is not sensible to calculate a mean, or are event counts, such as the number of asthma attacks per month. Although there are methods for dealing with these data, often the measures are dichotomised and treated as binary data.

A measure of the effect of an intervention is generated by comparing outcomes in the intervention group with those in the control group (see Box 7.3). The objective is to determine the extent to which outcomes are better or worse in the intervention group compared to the control group. Depending on the measurement scale of the outcome, an effect measure can be generated as a change in an event rate, or as a change on a continuous scale. For event data, this comparison could be generated in terms of relative differences (odds ratio and relative risk) or absolute differences (absolute risk reduction and number needed to treat) between the groups. For continuous data the effect measures are based on differences in means, or standardised mean differences (d-statistics, z-scores or effect sizes).

Box 7.3

Measures of effects of healthcare interventions

Effect measure (treatment effect, estimate of effect)

The observed relationship between an intervention and an outcome. This could be summarised as a p-value, odds ratio, relative risk, risk difference, number needed to treat, standardised mean difference, or weighted mean difference.

p-value (statistical significance)

The probability that the observed results in a study could have occurred by chance. A p-value of less than 5% (i.e. $p < 0.05$) is generally regarded as statistically significant.

Effect measures for binary data

- **Odds** The ratio of the number of people in a group with an event to the number without an event. Thus, if out of 100 people, 20 had the event (and 80 did not), and the odds would be 20/80 or 0.25.
- **Risk (proportion, probability or rate)** The proportion of participants in a group who are observed to have an event. Thus, if out of 100 patients, 20 had the event, the risk (rate of event) would be 20/100 or 0.20.
- **Odds ratio (OR)** The ratio of the odds of an event in the experimental (intervention) group to the odds of an event in the control group. An OR of one indicates no difference between comparison groups. For undesirable outcomes an OR that is less than one indicates that the intervention was effective in reducing the risk of that outcome.
- **Relative Risk (RR) (risk ratio, rate ratio)** The ratio of risk in the intervention group to the risk in the control group. An RR of one indicates no difference between comparison groups. For undesirable outcomes an RR that is less than one indicates that the intervention was effective in reducing the risk of that outcome.
- **Absolute risk reduction (ARR) (risk difference, rate difference)** The absolute difference in the event rate between two comparison groups. A risk difference of zero indicates no difference between comparison groups. For undesirable outcomes a risk difference that is less than zero indicates that the intervention was effective in reducing the risk of that outcome.
- **Number needed to treat (NNT)** The number of patients who need to be treated to prevent one undesirable outcome. It is the inverse of ARR.

Effect measures for continuous data

- **Mean difference** The difference between the means (i.e. the average values) of two groups.
- **Weighted mean difference (WMD)** Where studies have measured the outcome on the same scale (e.g. weight), the weight given to the mean difference in each study is usually equal to the inverse of the variance.
- **Standardised mean difference (SMD)** Where studies have measured an outcome using different scales (e.g. pain may be measured in a variety of ways) the mean difference may be divided by an estimate of the within-group standard deviation to produce a standardised value without any units.

Effect measure for survival data

- **Hazard ratio** A summary of the difference between two survival curves. It represents the overall reduction in the risk of death on treatment compared to control over the period of follow up of the patients.

It is important to remember that individual studies, because of small sample sizes, may not be able to estimate effects precisely.⁶ By combining the data from these studies, a meta-analysis acquires the statistical power to increase the precision of the estimate of effect. Although statistical significance is a prerequisite for confidence in the

precision of the results, it is not a measure of the magnitude of the effect.^{7, 8} Reviewers should be more interested in clinical importance of the results because it gives us the tools to make a judgement about the magnitude of the benefits or harm, and it helps in decision making.⁹⁻¹¹

It is well known that the manner of summarising results of randomised trials may influence judgements about patient management.¹²⁻¹⁵ Serious thought should be given to the choice of the effect measure that will be used to summarise the findings of the quantitative analysis. In particular, the odds ratio is criticised for not being well understood by physicians and patients.¹⁶ In this regard relative risk is more intuitively understandable. Whilst at lower baseline rates (<10%) the numerical value of odds ratio is similar to relative risk, if all odds ratio are misinterpreted as relative risks, the value of the interventions will be exaggerated, especially at higher baseline rates where odds ratios are always more extreme than relative risks.¹⁶

However, some statisticians prefer odds ratios to relative risks due to their mathematical properties.¹⁷ Unlike risk ratios and risk differences, odds ratios do not have inherent range limitations associated with high baseline rates.¹⁷ Moreover, the odds ratio naturally arises as the antilog of coefficients in mathematical modelling which makes it more suitable for statistical manipulation.¹⁷ However, the importance of these mathematical concerns is debatable. Mathematical properties of more importance include the availability of a variance estimator for the summary statistic (which prevents the NNT being used as a summary statistic) and the reversibility of the event classification. This latter point is an issue for risk ratios-in a meta-analysis of a mortality endpoint, analysis of risk ratios of survival will give a different answer to analysis of risk ratios of death. Treatment effects are described as consistent if they are relatively constant across different trials. Importantly, if there is variation in baseline event rates between the trials, relative measures of effect (the odds ratio and risk ratio) have been shown to be more likely to be consistent than absolute measures of effect (the risk difference).

Knowledge of the baseline event rate is also important for decision making in health care if treatment effects are assumed to work on relative scales. Effective treatments may only be recommended for use in high-risk groups where a high proportion of patients will benefit. The costs and morbidity associated with the intervention may outweigh the benefits in low risk patients.¹⁸⁻²⁰ So despite relative treatment effects being the most consistent summary measures, they may not communicate all the information required by decision-makers, who would prefer absolute effects. One approach to the problem is to use a relative summary statistic in the meta-analysis, but to express predicted benefits of treatment across a range of typical baseline event rates, either using risk differences (multiplication by 100 gives the number of events saved per 100 treated), or numbers needed to treat.^{9, 11, 21, 22} Given an efficacious treatment, the lower the number needed to treat, the smaller the number of patients needed to be treated to prevent/achieve a target outcome, and the higher is our confidence in the treatment. Conversely, the higher the number needed to treat, the greater the number of patients clinicians must treat to prevent/achieve a target outcome and the less inclined we would be to treat.

For data on a continuous scale, combining mean treatment effects (weighted mean differences) is straightforward when all measurements are comparable and on the same scale.²³ On other occasions it is necessary to transform the mean effect from each study to a standardised value before they can be pooled. The most common approach is to divide the difference in means by the sample standard deviation from each study- thus expressing the treatment effect relative to the natural variation expressed on the same scale. These standardised values (standardised mean differences) are commonly known as ‘d-statistics’ or effect sizes. Their use should be questioned when there may be other reasons why the observed variability of the outcomes within each study may not be constant.

One issue of concern in the meta-analysis of continuous data is the importance of skew. Many biomedical measurements, especially concentrations, have severely positively skewed distributions. The methods of meta-analysis will not be adversely effected by a mild degree of skew, but can give misleading results when skew is more severe. In such situations if log-transformed data are available (geometric means) they should be used in the analysis.

2.7.4.2 Methods of meta-analysis

Once the results of each study are summarised using an effect measure (e.g. odds ratio or relative risk or mean difference etc) an average value of the effect can be computed across the studies. One common misconception is that meta-analysis treats the data from the studies as if it all arose in a single study. This approach is naïve and the results can be misleading. Averaging summary statistics ensures that the intervention and control groups within each study are compared directly (participants in each study are only compared to other participants in the same study retaining the randomisation), and that the consistency of the results between the studies can be investigated.

Typically, the pooled effect estimate represents a weighted average of all studies included in the meta-analysis. The weights assigned to individual results are usually in inverse proportion to their variance, a method which gives more weight to the larger studies and less weight to the smaller studies.²⁴⁻²⁷ It is also possible to weight studies in relation to other factors like quality.²⁸

The pooling of results in a meta-analysis can be carried out using either a ‘fixed effect’ or a ‘random effects’ statistical model (see Box 7.4). A fixed effect model estimates the treatment effect as if there were a single ‘true’ value underlying all the study results. A random effects model, on the other hand, assumes that there is no single underlying value of effectiveness, but a distribution of values depending on known and unknown characteristics of the studies. In this case the differences between study results are considered to arise from the between study variation in underlying effectiveness and the play of chance. Whilst fixed effect methods simply estimate the average treatment effect and its confidence interval, random effects models estimate a mean treatment effect, its associated confidence interval, and the observed variance of the treatment effects between the studies (assuming that they have a normal distribution). One may view the fixed effect model as a special case of random effects model where between-study variation is taken to be zero.

There is disagreement as to which statistical model is most appropriate.²⁹⁻³¹ This is partly because of the differences in their inherent assumptions and partly due to the variation in pooled results when different methods are applied to the same data set. Random effects models weight smaller studies proportionally higher than fixed effect models, which in some circumstances may lead to different estimates. This phenomenon may exaggerate the impact of publication bias and poor quality, as smaller trials are the most likely to be affected by publication bias and also the most likely to have substandard methodology. Where there is heterogeneity between the trials the random effects model produces wider confidence intervals of the combined estimate compared to a fixed effect, such that the results are always more conservative. This is due to the fact that the random effects model takes into account both random variability and the variability in treatment effects between the studies which the fixed effect model ignores. Therefore, it can be argued that the fixed effect model may overestimate the precision of the treatment effect if there is significant unexplained heterogeneity between the studies. In practice both statistical models usually produce very similar results and the robustness of the quantitative synthesis can be examined if both are reported. However, when there is unexplained heterogeneity between studies, it may be more appropriate to use a random effects model as this approach accounts for the variation between studies that cannot be explained by other factors (see Phase 2.7.4.4).

For event data, commonly used fixed effect methods of combining odds ratios include the methods of Mantel-Haenszel³² and Peto.³³ The Mantel-Haenszel method yields reliable results in most situations, and has been extended to allow the combination of relative risks and risk differences.³⁴ The Peto method is simple to compute (it is in fact an approximation to the Mantel-Haenszel method), and it may be particularly useful for meta-analysis of binary data when event rates are very low where other methods fail.³⁵ The Peto method, however, does give biased answers in some circumstances, especially when treatment effects are very large, or where there is a lack of balance in treatment allocation within the individual studies.³⁶ This may be important when combining the results of observational studies which often have little balance. The standard calculations required to perform a random effects analysis have been described.³⁷ Logistic regression models can also be used to combine study results when the sample sizes are large.³⁸ This method also allows exploration of the influence of study characteristics, although when there is heterogeneity between studies it is preferable to use a random effects regression method.³⁹ The inverse variance method is a generic approach to pooling data, which, whilst rarely used for binary data outcomes, is commonly used for pooling differences in means and standardised effect sizes. The method can be extended to pool any summary statistic for which a standard error is known. Thus it can be used for pooling adjusted estimates, estimates corrected for clustering and repeat measurements, and other summaries derived from more complex statistical methods.

The analysis described so far is based on an approach to statistical analysis called the frequentist approach. An alternative approach, called Bayesian analysis, can be used in meta-analysis. It incorporates a prior probability distribution based on subjective opinion and objective evidence, such as the results of previous research. Using Bayes' theorem it updates the prior distribution in light of the results of the meta-analysis, producing a posterior distribution. Statistical inferences are based on this posterior distribution. The posterior distribution can also act as the prior distribution for new research. Bayesian models are closely related to random effects models,⁴⁰ and comprehensive methods for

Box 7.4

Key concepts in quantitative synthesis

Systematic error (bias)

A deviation in results either exaggerating or underestimating the 'true' value of the effect of an intervention. It can arise from systematic differences in the groups that are compared (selection bias), the care that is provided, or exposure to other factors apart from the intervention of interest (performance bias), withdrawals or exclusions of people entered into the study (attrition bias) or how outcomes are assessed (detection bias). The extent to which the design and conduct of a study are likely to produce or prevent systematic errors is evaluated by study quality assessment. More rigorously designed (better 'quality') studies are more likely to yield results that are closer to the 'truth'.

Random error (sampling error)

Error due to the play of chance that leads to imprecision in estimates of effect. Confidence intervals reflect the magnitude of random error.

Confidence interval (CI)

The range within which the 'true' value of the effect of an intervention is expected to lie with a given degree of certainty. Confidence intervals represent the distribution probability of random errors, but not systematic errors (bias).

Fixed effect model

A mathematical model for combining the results of studies that assumes that the effect is truly constant in all the populations studied. Thus, only within-study variation is taken to influence the uncertainty of results and it produces narrower confidence intervals than the random effects model.

Random effects model

A mathematical model for combining the results of studies that allows for variation in the effect amongst the populations studied. Thus, both within-study variation and between-studies variation are included in the assessment of the uncertainty of results.

performing Bayesian meta-analyses have been outlined.^{41, 42} These approaches have many attractive features, but they are controversial because they rely on opinions for generating prior probabilities which frequently vary considerably.

A selected list of available meta-analysis software is shown in Appendix A4.1.

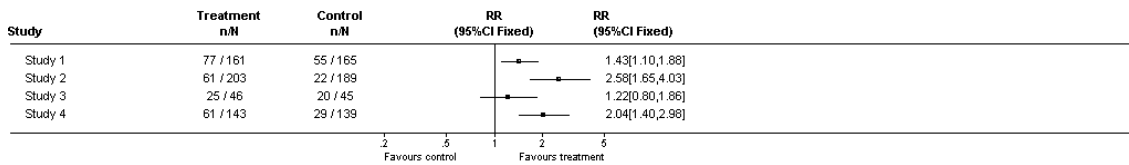
2.7.4.3 Presentation of quantitative results

Following a data synthesis, it is preferable to express the results in formats which are most easily comprehensible. It is possible to graphically display effect estimates and ranges of all the primary studies along with the overall summary, if appropriate on the same axis.⁴³ The most commonly used graphical approach is called the forest plot (see Box 7.5). It presents the individual study effects with their confidence intervals as horizontal lines, the box in the middle of the horizontal line representing the mean effect.⁴⁴ When using odds ratios or relative risk as the effect measure, the effects are usually plotted on a log-scale to introduce symmetry to the plot. The vertical line drawn at an odds ratio or relative risk of one (unity) represents 'no effect' and a confidence

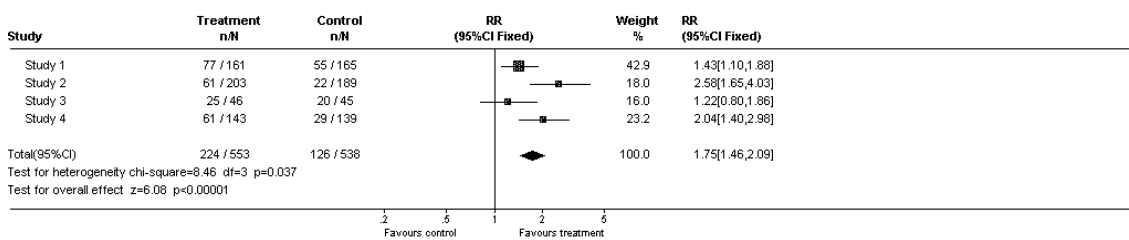
Box 7.5 Presentation of results

Effects of four trials included in a systematic review

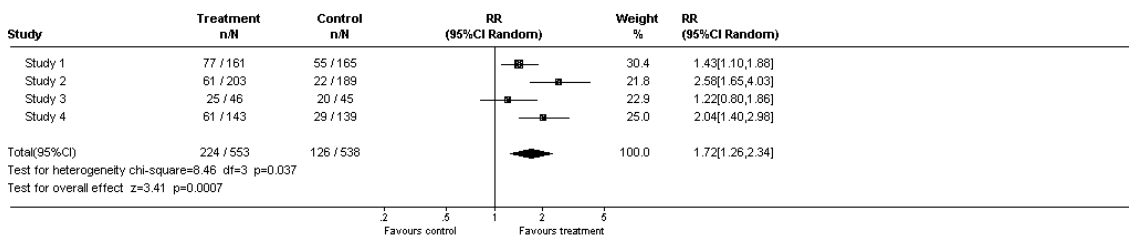
a) Presented without meta-analysis



b) Presented with meta-analysis (fixed effect model)



b) Presented with meta-analysis (random effects model)



interval overlapping this vertical line represents the lack of a statistically significant effect. An example of such a diagram is shown in Box 7.5. If risk difference is used as the effect measure, a vertical line at zero represents no effect. Different sized boxes may be plotted for each of the individual studies, the size of the box increasing with the weight that the study takes in the analysis. Other suggestions for graphical displays have also been proposed.⁴⁵

2.7.4.4 Investigating differences between studies

The variation or differences between estimates of effects in the component studies in a systematic review is described as heterogeneity. Homogeneity on the other hand is the degree to which the results of studies included in a review are similar. The investigation of differences in estimates of the treatment effects between studies included in a review is an essential part of data synthesis.^{1, 46-48}

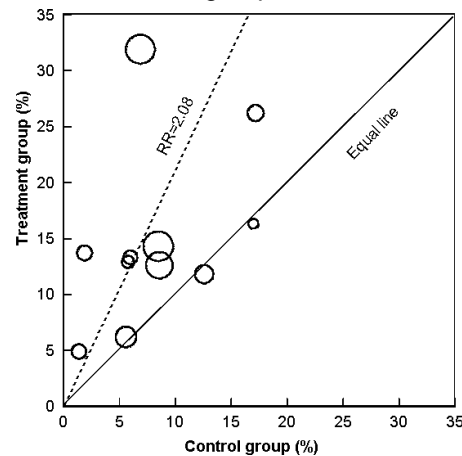
Box 7.6

Suggested steps in exploring for heterogeneity of results

- **Graphical exploration**
 - Forrest plot (see example in Box 7.5)
 - L'Abbé plot (see example below)
- **Statistical tests of homogeneity of 2x2 tables**
- **Stratified analysis or subset analysis**
- **Statistical regression modelling**

An example of graphical exploration using L'Abbé Plot (based on Song⁵⁰)

Plot of event rates in intervention and control groups of 11 trials in a systematic review⁵¹



Each circle represents an individual trial and the size of circles represents relative trial size. Solid diagonal line represents the line of equality of event rates in the two arms within trials. The dotted line which represents a summary RR of 2.08 may be called as 'the overall RR line', which is estimated by pooling the results of all eleven studies. It can be seen that the event rates vary greatly in both the treatment (4.9% to 31.9%) and the control (1.4% to 17.2%) groups indicating heterogeneity. Statistical tests of later show significant heterogeneity with Chi-square 35.26 (df=10) and $P < 0.001$.

At the simplest level, an idea of the heterogeneity of results between studies can be obtained by visually examining the forest plot of a meta-analysis for variations in effects (see Box 7.5). A useful adjunct to this graphical exploration is the use of L'Abbé plot, which plots event rates in the intervention group against event rates in the control group^{49, 50} (see Box 7.6). If the effects are homogeneous, the points would lie around a line parallel to the line of identity (equal line in the plot in Box 7.6); large deviations would indicate heterogeneity.^{50, 51} These graphical approaches should be used in conjunction with formal statistical tests to examine between-study differences.⁵² Statistical tests of homogeneity are used to assess whether the observed variability in study results (effect sizes) is compatible with that expected to occur by chance.⁵³ If the test of homogeneity is not significant it is still possible that there may be important between-study differences as these tests have low statistical power.⁵⁴

When statistically significant heterogeneity is detected or when such heterogeneity is suspected on graphical exploration, differences in the characteristics of the studies or other factors should be investigated as possible explanations. These differences can be summarised in a narrative synthesis, but where possible they should be formally estimated and evaluated. In stratified analysis, separate meta-analyses can be carried out in subsets of studies grouped according to one or more particular characteristics. The significance of differences in summary estimates in the subsets indicate heterogeneity. It may be possible to investigate the influence of differences in the participants and the intervention (e.g. age and baseline risk of the participants; dose and duration of the treatment), as well as differences in the definitions and measurement of outcomes. Meta-regression can be used to investigate the effects of differences in the study characteristics on the estimates of the treatment effect.^{38, 39, 55} These methods are best used when the characteristics under investigation are measured on a continuous scale. Effectively they fit linear regression models for each covariate, weighting each study according to the precision of the estimate of the treatment effect. Such analyses are best undertaken using random effects regression models.

The interpretation of the results of investigations into heterogeneity must be treated with some caution. Even where the original data have come from RCTs, the investigation of between-study differences is equivalent to an observational study investigating differences between sub-groups rather than differences between randomised groups.^{56, 57} On such occasions there may be other explanations for the observed differences. *A priori* comparisons which are planned in advance on the basis of a scientific theory and written into the protocol are more credible than findings which are found through *post hoc* exploratory analysis or data dredging - it is always possible to generate some explanation for the observed differences - but if multiple analyses have been undertaken it is most likely to be a spurious finding.⁵⁸ If the differences cannot be explained it has been argued that the synthesis should take the form of a random effects model and particular caution must be taken in the interpretation of the pooled estimate.³¹

One note of caution is the investigation of relationships between baseline risk and treatment effects. As the baseline risk and the treatment effect are calculated from the same numbers, they are naturally correlated. This can cause confusion in a meta-regression - it being difficult to ascertain whether there is a true relationship between risk group and effectiveness. Special methods of meta-regression have been developed for use in these situations.

2.7.4.5 Quantitative synthesis using individual patient data

Quantitative syntheses based on thoroughly checked and updated individual patient data are considered the most robust and reliable. This is because the reviewer can analyse all trials in a standard manner, dealing with drop-outs and missing values consistently, and including updated data not available from the individual reports of trials.^{59, 60} Reviews of this form require initiation and maintenance of an international collaborative group, and there are implications for resources and time.⁶¹ The main advantage is that analysis using individual patient data allows reviewers to undertake time-to-event (survival) analyses and subgroup analyses to test important hypotheses about differences in effect. In particular, when the outcome is time-dependent, e.g. in quantifying the effectiveness of cancer or infertility treatment, survival analyses are preferable to comparison of simple proportions.

2.7.4.6 Synthesis of studies with different designs

When studies of different designs are included in a systematic review, it is important that the potential biases that could be introduced by statistical combination are investigated.⁶² There are several ways in which an analysis can be undertaken to do this.²⁸ One approach is to separately synthesise the results of subgroups of studies with different designs or levels of validity and to compare the summary estimates of the subgroups for trends and important differences. An alternative approach is to cumulatively combine studies of decreasing strength of evidence and monitor changes in the overall estimates when studies of lower validity are included (a form of sensitivity analysis). Producing a plot of the study effects in decreasing order of validity may assist with this. A third approach involves modelling the strength of evidence as a variable in a regression analysis similar to that used in exploration of causes of heterogeneity. This requires each study to be given a grading according to its quality or validity. This method may be useful to describe systematic relationships between the validity of the primary studies and their results, and may give insights into the value of different methodological approaches.⁶³ However, there is no consensus on study quality scoring systems and so the result might vary according to the system used. When synthesising results of studies with different designs, non-quantitative synthesis is often the only feasible option.

2.7.4.7 Sensitivity analyses

Sensitivity analysis involves repeating the analysis whilst making some changes to the data.⁶⁴ Sensitivity analyses should be used to investigate how robust the overall findings of the review are to the review process, especially where there has been some uncertainty or disagreement involving inclusion of studies, data extraction, missing data and in the choice of statistical method. Redoing the analysis whilst changing each option will indicate how robust the review's conclusions are to these uncertainties. Sensitivity analyses can also be performed where there are one or more large studies which tend to dominate the results. A reanalysis excluding these studies will assess the degree to which they affect the review's conclusions.

The validity of a meta-analysis can be compromised by missing data and drop-outs in individual studies. The review should report the number of participants who are included in the final analysis as a proportion of all participants in all studies. If the number of participants missing from the final analysis is large it will be informative to detail the reasons for their exclusion. In some circumstances it may be possible to estimate missing data.⁶⁵ Alternatively the degree to which the review's conclusions could be altered by the missing data should be investigated in a sensitivity analysis by alternately substituting the least favourable and most favourable outcomes in place of the missing data in best case and worst case scenario analyses.

2.7.5 Data synthesis of test accuracy studies

Quantitative synthesis of trials assessing the impact of a testing strategy (test and therapeutic interventions based on the test result) on outcome is performed as described above (see Phase 2.7.4). This can only be done when there is consensus about treatment strategies, as is often the case for screening interventions. When conditions for which effective interventions are known to exist, it may be appropriate to assess the accuracy of tests without the need to conduct randomised trials.⁶⁶ However, quantitative synthesis of studies evaluating accuracy of diagnostic tests is

Box 7.7

Measures of diagnostic accuracy of dichotomous tests

Sensitivity (true positive rate)

The proportion of those people who really have the disease who are correctly identified as having the disease.

Specificity (true negative rate)

The proportion of those people who really do not have the disease who are correctly identified without the disease.

Positive predictive value

The proportion of the people who test positive who truly have the disease.

Negative predictive value

The proportion of the people who test negative who truly do not have the disease.

Likelihood ratios (LRs)

The LR is the ratio of the probability of a positive (or negative) test result in patients with disease to the probability of the same test result in patients without disease. The LR indicates by how much a given test result raises or lowers the probability of having disease. With a positive test result, a LR >1 increases the probability that disease will be present. The greater the LR, the larger the increase in probability of the disease and the more clinically useful the test result. With a negative test result, a LR <1 decreases the probability that disease is present: the smaller the LR, the larger the decrease in the probability of disease and the more clinically useful the test result.

Diagnostic odds ratio

The ratio of likelihood ratios. Provides a single measure of accuracy.

Summary receiver operating characteristics curve (SROC)

Summarising the performance of a dichotomous test.

somewhat different to that used for studies of effectiveness. The main principles and methods are described below.

2.7.5.1 Choice of accuracy measure

There are three sets of summary statistics which are commonly used to report diagnostic accuracy. Sensitivity and specificity describe the ability of a test to correctly identify individuals with the disease and individuals without the disease respectively, predictive values give the probabilities that positive and negative test results are actually correct, whilst likelihood ratios describe the relative chances of obtaining each test result in individuals with and without disease. All these measures are paired: single measures of accuracy are seldom used in primary studies. The powerful properties of likelihood ratios (LRs) lead them to be preferred over sensitivity and specificity in many clinical situations, and can clearly demonstrate the clinical value of a test.⁶⁷⁻⁷⁰ Misleading inferences concerning the value of diagnostic tests may be made without the use of LRs.⁷¹

Test accuracy is however most commonly reported using sensitivity and specificity. The results of a set of studies can be displayed on a receiver operating characteristics (ROC) plot, plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). Variability between the studies will be seen in both sensitivity and specificity, due to sampling variation and between study variations. In particular, if the studies differ in the threshold used to define a positive test result, there will be a particular pattern noted – studies which have higher sensitivity will tend to have lower specificity, whilst those with lower sensitivity will have higher specificity.⁷² There is also a similar relationship between positive and negative likelihood ratios. In this situation these measures do not behave independently, and it does not make sense to generate separate averages for sensitivity and specificity simply by pooling their results from different studies.⁷³

2.7.5.2 *Methods of meta-analysis*

There is considerable debate about how best to statistically summarise results from several test accuracy studies.^{66, 74, 75} There are four commonly used approaches: averaging sensitivities and specificities, averaging positive and negative likelihood ratios, pooling diagnostic odds ratios and fitting summary ROC curves. The first two methods have the benefit of providing clinically interpretable statistical summaries, but are inappropriate methods to use when there may be variation in the test threshold between the studies. In such situations the average sensitivity and specificity (or likelihood ratios) will be an underestimate of the performance of the test.

When there is a threshold effect, the data are best summarised as an ROC curve, known as a summary receiver operating characteristic (SROC) curve.⁷⁶⁻⁷⁸ One approach to calculating such a curve involves re-expressing the treatment effects as diagnostic odds ratios that summarise the diagnostic capabilities of the test as a single statistic. Even when there is variation in thresholds, it is possible that the diagnostic odds ratios are constant or nearly constant between studies, which correspond to a pattern of symmetrical lines on the SROC plot. Alternatively non-symmetrical SROC lines (where even the diagnostic odds ratio vary with the threshold) can be estimated⁷⁹ which involve regression between the sums and differences of logistic transformations of the true positive and false positive rates. Whilst the SROC method copes with variations in test thresholds, it is difficult to interpret clinically.

As there is a divergence of opinions about the most suitable approach for meta-analysis of test accuracy studies, reviewers may wish to use several approaches to determine if their summary results are sensitive to the variation in statistical methods used.

2.7.6 **Publication bias**

The accessibility of research results is dependent not only on whether a study is published, but also on when, where and in what format. The level of accessibility is related to several factors including the selective publication of statistically significant results, the timing of publication, the type and language of publication, multiple publications, selective citation of references, and coverage by database and indexes.⁸⁰⁻

⁸⁴ All of these factors influence the identification of relevant studies and without the use of systematic approaches to track down less accessible studies, reviews can become

biased. Systematic reviews should attempt to overcome these biases by using search strategies that include a variety of searching methods (both computerised and manual) and explore multiple overlapping sources of research evidence (see Phase 2.3.2).

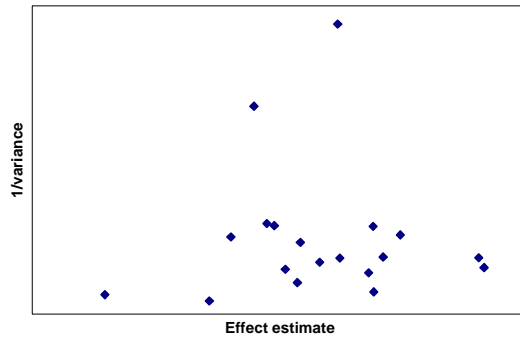
Because even thorough literature searches cannot eliminate the risk of publication bias, formal assessment and estimation of such a risk should be incorporated into the review's analysis, conclusions and inferences.⁸⁴⁻⁸⁶ A range of statistical and modelling methods are available to deal with publication bias in meta-analyses, such as the fail-safe N (or file drawer) method,⁸⁷ funnel plot,⁸⁸ rank correlation method,⁸⁹ linear regression method,⁹⁰ trim and fill method,^{91, 92} and some complex modelling methods.⁹³ Statistical methods based on analysis of funnel plot symmetry are most commonly used. Funnel plots show the distribution of effect sizes according to sample size (or inverse of variance): it is to be expected that the points will fill a funnel shape, there being more variability in reported effect sizes for smaller studies (see Box 7.8). Large gaps in the funnel indicate a group of possibly 'missing' publications. These omissions are usually small studies with point estimates suggesting a different effect from those available and are unlikely to be missing at random. However, there is a need for caution in interpretation as the shape of a funnel plot is dependent on the measures selected for estimating effect and precision.⁹⁴

These methods are mainly used to detect the possibility of publication bias, but some methods (for example, the trim and fill method) could provide an estimate by adjusting for the publication bias assumed detected.⁹³ All these methods are by nature indirect and exploratory, because the true extent of publication bias is generally unknown. In addition, there are some methodological difficulties in using the available methods to assess publication bias in meta-analyses. For example, in most cases, it is impossible to separate the influence of factors other than publication bias on the observed association between the estimated effects and sample sizes across studies. Moreover, the appropriateness of many methods is based on some strict assumptions that can be difficult to justify in practice. For these reasons, it seems reasonable to argue that these methods are not very good remedies for publication bias.⁸⁰ The attempt at identifying or adjusting for publication bias in a meta-analysis should be mainly used for the purpose of sensitivity analyses, and the results should be interpreted with caution. Increasingly, with the growth in registration of controlled trials before their results are known (see Phase 2.3.6), it will become possible to do sensitivity analyses to assess whether analyses restricted to such trials differ from those using data from all trials.⁹⁵

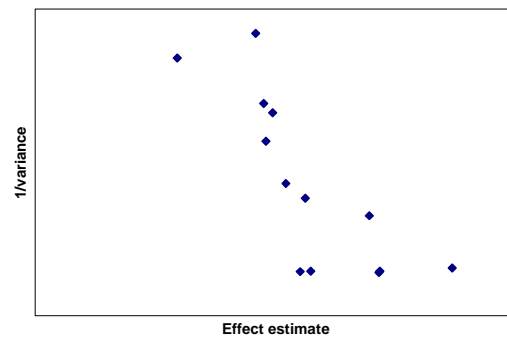
Box 7.8 Examples of funnel plots and reasons for asymmetry

Examples of funnel plots

Symmetrical



Asymmetrical



Reasons for funnel plot asymmetry

- Publication bias
- Location biases
- English language bias
- Database bias
- Citation bias
- Multiple publication bias
- Bias in provision of data
- Poor methodological quality of small studies
- Clinical heterogeneity e.g. small studies in high risk populations

2.7.7 Data synthesis of qualitative research

Findings from studies using qualitative methods can help to inform the various features of data synthesis such as exploration of the diversity of effects across studies, settings and groups; and investigation of average and divergent effects.⁹⁶ They can, for example, identify differences between interventions in apparently similar studies or between the context within which interventions are delivered. They may also illuminate the impact of contextual factors, such as the qualitative impact of unexpected events in the delivery of the intervention. An example of synthesis of findings from qualitative research in an effectiveness review is given in Box 7.9.

Data synthesis of qualitative research is far from simple. In theory at least, the synthesis of findings from qualitative studies may be narrative or meta-analytical. To date, systematic reviews of effectiveness that have sought to incorporate qualitative research have usually included the data synthesis in the discussion section. There are no formal procedures available to aid narrative synthesis of findings from qualitative studies within the context of a systematic review. However, the same criteria used to judge the quality of the studies to be included in the synthesis could be applied to the synthesis itself. In particular, it is important that the process whereby conclusions are

Box 7.9

An example of synthesis of research including qualitative data

Review Question

What is the effectiveness and appropriateness of peer-delivered health promotion for young people?

Population	Young people aged 11 to 24 years old.
Interventions	Peer-delivered health promotion initiatives (e.g. skill development, awareness raising or other approaches) across a range of health topics including drugs, smoking, safe sex and diet. Studies could compare peer-delivered interventions to those delivered by adults or to groups receiving no intervention.
Outcomes	Effectiveness: A range of outcomes including health related behaviours, attitudes, intentions and knowledge. Appropriateness: A range of process outcomes including acceptability and accessibility of the intervention, factors influencing the implementation of the intervention, collaborations and partnerships, and delivery of the intervention.
Study designs	Effectiveness: Randomised and non-randomised studies using quantitative methods. Appropriateness: Process evaluation studies using both quantitative and qualitative methods.

Synthesis of results from effectiveness studies

There were 12 effectiveness studies which were judged to be of sound quality. Non quantitative synthesis showed that more studies demonstrated peer-delivered health promotion to be effective than ineffective. However, it was not possible to identify specific characteristics of an effective model of peer-delivered health promotion.

Synthesis of results from process evaluation studies

There were 15 process evaluations of which two were embedded within effectiveness studies. The process evaluations included a range of designs and only two were assessed to be of sound quality (see Box 5.11). Synthesis of findings identified the following factors/processes that may influence the outcome of peer-led interventions:

Acceptability: Young people are more likely to express a preference for peer-led sessions in comparisons to teacher led and to perceive peer leaders as credible sources of information. There were few negative reactions to peer-led interventions but some studies reported young men being more uncomfortable than young women with emphasis on feelings.

Factors influencing implementation: Many factors were identified. Conflict between the philosophy of peer education as a non-traditional educational strategy implemented in more traditional school settings was particularly common. In these settings teachers could undermine peer leaders.

Training and personal development of peer leaders: On going support for peer educators was identified as a high priority. Peer leaders also reported a range of positive personal development outcomes linked to their training and experience but the studies were not designed to assess change in these outcomes in a reliable way.

Accessibility: Only limited data were available on this process. However, the review suggests that young people acting as peer educators might be most comfortable working with their friends and even in this context may feel that the advice and/or information they were to transmit would be seen as 'interfering'.

Recruitment of peer leaders: A consistent finding from the process evaluations was that peer leaders were more likely to be female and that once recruited it was more difficult to retain male peer leaders. Some interventions did succeed in recruiting peer leaders from 'at risk' groups but others were more likely to recruit high achievers.

Working in partnership with young people: The studies reviewed highlighted the potential for conflict between adult co-ordinators of interventions and the peer leaders they were supporting. In particular, adults involved may find it difficult to deal with the increasing confidence of peer leaders and their emerging ideas, wants and needs. When things do not go according to previously agreed plans, there is a tendency for professional co-ordinators to try to regain control and this can undermine the peer leaders.

Based on a review by Harden et al¹¹⁶

drawn from study findings are made as transparent as possible and that there are attempts made to replicate conclusions.

Analytic synthesis of qualitative research findings can take a number of different approaches. Clearly, the feasibility of attempts will depend on the diversity of the studies involved. It is highly unlikely that such a synthesis will involve a re-analysis of primary data which may be in the form of transcripts from interviews, or field-notes from studies involving participant observation. Rather, the data to be analysed are most likely to be the findings of the studies involved. These might take the form of substantive themes arising, for example, from in-depth interviews. Within qualitative research (and arguably all research) theory plays a pivotal role in informing the interpretation of data. Whilst few authors appear to have considered the role for theory-led synthesis of findings across studies an argument can be made for exploring the potential for this approach.⁹⁷

In terms of analytical approaches there are parallels between synthesis across studies and analysis associated with multi-site or multi-case research.^{98, 99} Qualitative research involving comparison between sites/cases can increase confidence in the generality of a finding or explanation.¹⁰⁰ Most of the techniques developed to aid the synthesis of qualitative data across sites/cases assume that data have been collected within a single study. This is the case with the meta-matrices proposed, for instance, by Miles and Huberman.⁹⁸ Noblit and Hare do attempt to grapple with the intricacies of meta-analysing ethnographic data from different sites and studies.¹⁰¹ As with data synthesis in quantitative research, there is a need for careful assessment of the comparability of the data to be synthesised.

A related issue is the potential for combining qualitative and quantitative findings from different studies. One approach to this type of synthesis is being developed using Bayesian hierarchical modelling.^{102, 103} Triangulation may provide another approach. This is a widely accepted technique for exploring the validity of, and relationship between, findings from research through the systematic comparison of data collected from different perspectives. Although normally involving comparisons of qualitative and quantitative data collected during the same study this approach can be applied to data from different studies.

There are, inevitably, objections to triangulation in particular and attempts to combine findings from qualitative and quantitative studies in general. Ackroyd and Hughes, for example, question the epistemological validity of the triangulation of qualitative and quantitative data, arguing that there are no 'criteria upon which all agree and which can be used to decide between alternative theories, methods and inconsistencies in findings'.¹⁰⁴ In contrast, Mason makes a strong case for the synthesis of different types of data, arguing that the key question is 'what are all the components necessary for generating a viable and convincing explanation and how do we get to that point'.¹⁰⁵

Another approach to the synthesis of information from different types of sources, which may be relevant to the review of findings from qualitative research, underpins consensus methods, such as the Delphi process and the nominal group technique. These approaches allow a wider range of study types to be considered than is possible

in a statistical meta-analysis, by focusing on the resolution of inconsistencies in results and attempting to provide quantitative estimates of expert opinions on results from qualitative processes.¹⁰⁶

2.7.8 Data synthesis of economic evaluations

The methods for syntheses of economic evaluations are not as advanced as those used for clinical effectiveness.¹⁰⁷⁻¹⁰⁹ The aim of data synthesis of economic evaluations is to summarise the evidence about the efficiency of health care provision to reduce the uncertainty about relative benefits and costs associated with alternative interventions. As indicated earlier, the first step should be to evaluate the evidence about clinical effects. If this review shows reliable evidence of equivalent or increased effectiveness, the next step will be to assess the available economic evidence. Where the results of clinical evidence used in economic evaluations are at odds with the findings of the effectiveness review, this would warrant further investigation. Therefore, when comparing and summarising findings of economic evaluations, it is important to remember that greater weight should be given to those good quality studies that are less subject to bias (see Phase 2.5.8).

In data synthesis of economic evidence, the initial analysis is usually non-quantitative. First, the included studies are described in terms of type of evaluation, setting and location, and perspective (hospital, health service, society, third party payer, patient, etc.). Then the source of evidence (trial based or review/model based), approach to data collection and analysis (prospective resource use, source of cost data and outcomes, mixed prospective and retrospective, and retrospective), and features of study quality (see Phase 2.5.8) are considered. Following this, the clinical and economic findings of the evaluations are summarised. A suggested format is given in Appendix 3.3. The results may be tabulated under several headings including quality of effectiveness evidence (experimental, observational, mixed, etc), magnitude of effectiveness (benefits), sources of cost data (hospital/patient records, reimbursement tariffs, literature review, etc.), and costs (direct and indirect). If the time frame of the analysis warrants discounting (>2 years), this should be reported if undertaken. If discounting is relevant but not undertaken, this should also be reported as a deficiency. To facilitate comparison between countries and time periods, it may be possible to standardise cost data by inflating or deflating cost to one specific year if sufficient details of economic analyses have been reported. A synthesis of costs and benefits based on the review (average and/or incremental cost-effectiveness ratios) can then be generated. When there is variability in the reliability of the data, the impact of appropriate sensitivity analyses on the estimate of cost-effectiveness should be taken into consideration.

Box 7.10

Possible permutations for results of economic evaluations¹¹¹

		Health outcomes		
		+	o	-
Costs	+	A	B	C
	o	D	E	F
	-	G	H	J

Legend

	Comparison	Health outcomes	Costs
+	intervention vs comparator	Better	Higher
o	intervention vs comparator	Same	Same
-	intervention vs comparator	Poorer	Lower

Implication of findings

A	Trade off	Higher costs but better outcomes (incremental analysis required)
B	Reject	Higher costs and no difference in outcomes
C	Reject	Higher costs and poorer outcomes
D	Accept	No difference in costs and improved outcomes (partial dominance)
E	Neutral	No difference in costs and no difference in outcomes
F	Reject	No difference in costs and poorer outcomes
G	Accept	Lower costs and improved outcomes (extended dominance)
H	Accept	Lower costs and no difference in outcomes (partial dominance)
J	Trade off	Lower costs but poorer outcomes (incremental analysis required)

Heterogeneity between populations, interventions, sources of data and methods of analysis usually means that a quantitative synthesis cannot be conducted.¹¹⁰ However, considering the tabulated information and the quality of the economic evidence, a conclusion about the direction and magnitude of cost-effectiveness can be reached for each one of the included evaluations¹¹¹⁻¹¹³ (see Box 7.11). For example, in terms of magnitude and direction, the findings may be that the intervention in question always dominates the comparator (more effective and less costly), in which case the review can conclude that both effectiveness and economic arguments support the use of the intervention¹¹⁴ (see Box 7.11). If such dominance is not demonstrated, then incremental benefit (incremental cost-effectiveness or incremental cost-utility) may be recorded. For example, if the intervention is more effective but also more costly, the review team would wish to examine differences in the magnitude of the decision in terms of incremental costs per additional unit of benefit gained. If magnitudes of incremental cost-effectiveness are similar then it will be possible to estimate the additional costs at a policy level (the payer's perspective). The reviewers will also need to be alert to the reporting of average cost-effectiveness ratios, which are misleading to the decision-maker.¹¹⁵ For example, if an intervention costs £100 and produces 10 units of benefit, its average cost-effectiveness is £100/10 = £10 per unit of benefit. An alternative intervention may cost £1000 but produce 50 units of benefit, in which case the average cost-effectiveness would be £1000/50 = £20 per unit of benefit. Faced with this information the decision-maker is led to believe that the first intervention is more cost-effective. However, what the decision-maker would find most useful is the incremental cost-effectiveness which equals (1000-100)/(50-10) = £22.50 per additional unit of benefit (the incremental cost-effectiveness). A

Box 7.11**An example of non-quantitative synthesis of economic evaluations****Findings of individual studies****Study no. Result**

	Costs	Outcomes
1	Lower costs	No difference in outcomes
2	Lower costs	No difference in outcomes
3	Lower costs	No difference in outcomes
4	Lower costs	No difference in outcomes
5	Lower costs	No difference in outcomes
6	Higher costs	Better outcomes
7	Higher costs	Better outcomes
8	No difference in costs	Improved outcomes
9	No difference in costs	Improved outcomes
10	No difference in costs	Improved outcomes
11	Lower costs and	Improved outcomes
12	Lower costs and	Improved outcomes

Permeation plot of results of individual studies

		Health outcomes		
		+	o	-
Costs	+	2	0	0
	o	3	0	0
	-	2	5	0

Narrative summary

The summary finding from these 12 studies is that the intervention is as/more effective and as/ less expensive as the comparator.

Based on a review of economic evaluations in community based care by Browne et al¹⁴

narrative overview of the efficiency estimates of each individual study can give a reasonable idea about the overall level of cost-effectiveness.

If the results of the non-quantitative synthesis show that there is residual uncertainty about efficiency, a new economic model can be built using the information collated in the review to produce a more robust estimate of cost-effectiveness.¹¹² Ways in which a meta-analysis of the findings of economic evaluations can be undertaken have also been reported.¹¹⁰ The standardised measure of economic value of an intervention (e.g. cost-effectiveness or cost-utility ratios) may be pooled across studies. However, meta-analysis requires comparability of the studies in terms of the methods used to assess the clinical effectiveness of the intervention (e.g. classification systems and preference weights) as well as homogeneity of the methods used to carry out the cost analysis. Moreover, this quantitative synthesis requires information about the sampling distribution of the data, including the cost-effectiveness ratios, and the results of statistical tests, which is often lacking in economic evaluations. For undertaking modelling or meta-analysis of economic data, expert input from health economists will be required.

2.7.9 Key points about data synthesis

- The aim of data synthesis is to collate and summarise the data extracted from primary studies included in the review.
- Data synthesis is the tabulation of study characteristics and results (non-quantitative synthesis), and use of statistical methods if appropriate (quantitative synthesis).
- Non-quantitative synthesis of evidence helps in planning quantitative syntheses and in some circumstances it may allow reviewers to qualitatively determine if the intervention of interest is likely to be effective, and if so under what circumstances.
- Quantitative synthesis focuses on use of statistical techniques to pool results from primary studies (meta-analysis), to evaluate heterogeneity of results, to assess for publication bias, etc.
- Findings from qualitative research may aid the interpretation of the quantitative findings.
- Economic outcomes and effects of interventions may be summarised using tabulated forms. Where there is residual uncertainty about cost-effectiveness, an economic evaluation may be modelled using parameters derived from the review.

2.7.10 References

1. Clarke M, Oxman A, editors. Section 8. *Analysing and presenting results*. Cochrane Collaboration, 2000. [cited 2000 December]. Available from: URL: http://www.cochrane.dk/cochrane/handbook/hbook8_ANALYSING_AND_PRESENTING_RESUL.htm
2. Hedges LV, Raudenbush SW, Hunter JE, Schmidt FL. Statistically analysing effect sizes. In: Cooper H, Hedges LV, editors. *Handbook of research synthesis*. New York, NY: Russell Sage Foundation; 1994. pp. 295-336.
3. Light RJ, Pillemer DB. In: *Summing up: The science of reviewing research*. Cambridge, Mass.: Harvard University Press; 1984. pp. 50-103.
4. Marcus SH, Grover PL, Revicki DA. The method of information synthesis and its use in the assessment of health care technology. *Int J Technol Assess Health Care* 1987;3:497-508.
5. Slavin B. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995;48:9-18.
6. Moye LA. P-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol* 1998;8:351-7.
7. Thompson WD. Statistical criteria in interpretation of epidemiologic data. *Am J Public Health* 1987;77:191-194.
8. Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ* 1994;309:727-730.
9. Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *Can Med Assoc J* 1995;152:351-7.
10. Guyatt GH, Cook DJ, Jaeschke R. How should clinicians use results of randomised trials? *ACP J Club* 1995;122:A12-A13.
11. Laupacis AL, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728-1733.
12. Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. *BMJ* 1995;311:1056-9.
13. Forrow L, Taylor WC, Arnold RM. Absolutely relative: How research results are summarised can affect treatment decisions. *Am J Med* 1992;92:121-4.
14. Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med* 1992;117:916-21.
15. Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *BMJ* 1994;309:761-4.
16. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol* 1994;47:881-889.
17. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987;125:761-768.
18. Smith GD, Egger N. Who benefits from medical interventions. *BMJ* 1993;308:72-74.
19. Guyatt GH, Sackett DL, Sinclair J, Hayward R, Cook DJ, Cook R. Users' guide to the medical literature. IX. A method for grading health care recommendations. *JAMA* 1995;274:1800-1804.

20. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995;311:1356-1359.
21. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310:452-454.
22. Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. *BMJ* 1996;312:426-429.
23. Wolf FM. *Meta-analysis: quantitative methods for research synthesis*. Beverly Hills, CA: Sage Publications; 1986.
24. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care* 1990;6:5-30.
25. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. San Diego, CA: Academic Press; 1985.
26. Fleiss JL. The statistical basis of meta-analysis. *Stat Methods Med Res* 1993;2:121-145.
27. Emerson JD. Combining estimates of the odds ratio: the state of the art. *Stat Methods Med Res* 1994;3:157-178.
28. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta analysis. *J Clin Epidemiol* 1992;45:255-265.
29. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;8:141-151.
30. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technol Assess* 1998;2. Available from: URL: <http://www.hta.nhsweb.nhs.uk/>
31. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991;338:1127-1130.
32. Breslow NE, Day NE. Combination of results from a series of 2x2 tables; control of confounding. In: *Statistical methods in cancer research, Volume 1: The analysis of case-control data*. Lyon: International Agency for Research on Cancer; 1980.
33. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: overview of the randomized trials. *Prog Cardiovasc Dis* 1985;27:335-371.
34. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985;41:55-68.
35. Deeks J, Bradburn M, Bilker W, Localio R, Berlin J. Much ado about nothing: statistical methods for meta-analysis with rare events [conference presentation]. In: 6th Cochrane Colloquium; 1998; Baltimore. Providence: New England Cochrane Center Providence Office. Available from: URL: <http://www.cochrane.org/colloquium/ps01.htm#abstract>
36. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Stat Med* 1990;9:247-252.
37. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-188.
38. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;10:1655-1677.
39. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995;14:395-411.

40. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to meta-analysis: a comparative study. *Stat Med* 1995;14:2685-2699.
41. Eddy DM, Hasselblad V, Schachter R. *Meta-analysis by the confidence-profile approach*. San Diego, CA: Academic Press; 1992.
42. Carlin JB. Meta-analysis for 2x2 tables: a Bayesian approach. *Stat Med* 1992;11:141-158.
43. Light RJ, Singer JD, Willet JB. The visual presentation and interpretation of meta-analyses. In: Cooper H, Hedges LV, editors. *Handbook of research synthesis*. New York, NY: Russell Sage Foundation; 1994. pp. 439-453.
44. Demets DL. Methods for combining randomized clinical trials: strengths and limitations. *Stat Med* 1987;6:341-348.
45. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889-894.
46. Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? *Stat Med* 1987;6:351-358.
47. Brand R. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1992;11:2077-2082.
48. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;309:1351-1355.
49. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;107:224-233.
50. Song F. Exploring heterogeneity in meta-analysis: is the L'Abbe plot useful? *J Clin Epidemiol* 1999;52:725-730.
51. Dolan-Mullen P, Ramirez G, Groff J. A meta-analysis of randomized trials of prenatal smoking cessation interventions. *Am J Obstet Gynecol* 1994;171:1328-1334.
52. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Heterogeneity. In: *Systematic reviews of trials and other studies*. 1998:39-54.
53. Paul SR, Donner A. A comparison of tests of homogeneity of odds ratios in k 2x2 tables. *Stat Med* 1989;8:1455-1468.
54. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios. *Stat Med* 1992;11:159-165.
55. Greenland S, Longnecker MD. Methods for trend estimation from summarised dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992;125:1301-1309.
56. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized controlled trials. *JAMA* 1991;266:93-98.
57. Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
58. Anello C, Fleiss JL. Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol* 1995;48:109-116.
59. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 1993;341:418-422.

60. Clarke MJ, Stewart LA. Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? *BMJ* 1994;309:1007-1010.
61. Stewart LA, Clarke M. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Stat Med* 1995;14:2057-79.
62. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;44:127-139.
63. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609-613.
64. Greenhouse JB, Iyengar S. Sensitivity analysis and diagnostics. In: Cooper H, Hedges LV, editors. *Handbook of research synthesis*. New York, NY: Russell Sage Foundation; 1994. pp. 383-398.
65. Pigott TD. Methods for handling missing data in research synthesis. In: Cooper H, Hedges LV, editors. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation; 1994. pp. 163-175.
66. Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests. *Screening and diagnostic tests: Recommended methods [online]*. Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests, 1996. [cited 2000 November]. Available from: URL: <http://hiru.mcmaster.ca/cochrane/cochrane/Sadtdoc1.htm>
67. Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? B. What are the results and will they help me in caring for my patients? *JAMA* 1994;271:389-391, 703-707.
68. Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA* 1994;271:703-707.
69. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Critically appraising the evidence. Is the evidence about a diagnostic test important. In: *Evidence-based medicine: How to practice and teach EBM*. London: Churchill Livingstone; 1997. pp. 118-28.
70. How to read clinical journals. II. To learn about a diagnostic test. *Can Med Assoc J* 1981;124:703-10.
71. Khan KS, Khan SF, Nwosu CR, Chien PFW. Misleading authors' inferences in obstetric diagnostic test literature. *Am J Obstet Gynecol* 1999;181:112-5.
72. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167-178.
73. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;2:S37-S47.
74. Irwig L, Tostenson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for the meta-analysis of diagnostic tests. *Ann Intern Med* 1994;120:667-676.
75. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-130.
76. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;13:1293-1316.
77. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarising diagnostic test performance: receiver-operating-characteristic-summary point estimates. *Med Decis Making* 1993;13:253-257.

78. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;13:313-321.
79. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic data. *Acad Radiol* 1995;2:S48-S56.
80. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990;163:1385-1389.
81. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992;263:374-378.
82. Grégoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? *J Clin Epidemiol* 1995;48:159-163.
83. Easterbrook P, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-872.
84. Song F, Eastwood A, Gilbody S, Duley L, Sutton A. Publication and related biases. *Health Technol Assess* 1999;4.
85. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc Ser A* 1988;151:419-463.
86. Begg CB. Publication bias. In: Cooper H, Hedges LV, editors. *Handbook of research synthesis*. New York, NY: Russell Sage Foundation; 1994. pp. 399-409.
87. Rosenthal R. The "file drawer problem" and tolerance for null results. *Psychol Bull* 1979;86:638-641.
88. Light RJ, Pillemer DB. *Summing up: the science of reviewing research*. Cambridge, Mass.: Harvard University Press; 1984.
89. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;50:1088-1101.
90. Egger M, Davey-Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-634.
91. Taylor S, Tweedie R. *A non-parametric "trim and fill" method of assessing publication bias in meta-analysis*. Technical Report: Colorado State University, Department of Statistics, 1998.
92. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. An empirical assessment of the impact of publication bias on meta-analyses. *BMJ* 2000;320:1574-1577.
93. Sutton AJ, Song F, Gilbody SM, Abrams KR. Modelling publication bias in meta-analysis: a review. *Stat Methods Med Res* 2001:In press.
94. Tang J-L, Liu J. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol* 2000;53:477-484.
95. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986;4:1529-1541.
96. Popay J, Williams G. Qualitative research and evidence based healthcare. *J R Soc Med* 1998;91.
97. Williams F, Popay J, Oakley A. *Welfare research: A critical review*. London: UCL Press; 1999.
98. Miles M, Huberman AM. *Qualitative data analysis: A source book of new methods*. Beverley Hills, CA.: Sage Publications; 1984.

99. Yin RK. *Case study research: design and methods*. 2nd ed. Thousand Oaks, CA: Sage; 1994.
100. Glaser BG, Strauss AL. *The discovery of grounded theory*. Chicago: Aldine; 1967.
101. Noblit G, Hare D. *Meta ethnography: synthesising qualitative data*. Newbury Park, CA: Sage; 1988.
102. Lilford R, Braunholtz D, Chard J. Reconciling the quantitative and qualitative traditions: the bayesian approach [conference presentation]. In: 7th Cochrane Colloquium; 1999; Rome.
103. Jones D, Abrams K, Dixon-Woods M, Roberts K, Sutton A. *Developing meta-analysis for health services research contexts*. Trent Institute Newsletter 1998;3.
104. Ackroyd S, Hughes J. *Data collection in context*. 2nd ed. London: Longman; 1992.
105. Mason J. Linking qualitative and quantitative data. In: Bryman A, Burgess RG, editors. *Analyzing qualitative data*. London: Routledge; 1994.
106. Jones J, Hunter D. Consensus methods for medical and health services research. In: Mays N, Pope C, editors. *Qualitative research in health care*. London: BMJ Publishing Group; 1996. pp. 46-58.
107. Cochrane Economics Methods Group. Cochrane Economic Methods Group. In: *The Cochrane Library [cd-rom]*. Issue 3: Oxford: Update Software; 2000.
108. NHS Centre for Reviews and Dissemination. *Making cost-effectiveness information accessible: the NHS Economic Evaluation Database project*. CRD guidance for reporting critical summaries of economic evaluations. York: University of York, NHS Centre for Reviews and Dissemination; 1996. Report No.: CRD report 6. Available from: URL: <http://www.york.ac.uk/inst/crd/crdrep.htm>
109. Jefferson T, Demicheli V. Methodological quality of economic modelling studies. A case study with hepatitis B vaccines. *Pharmacoeconomics* 1998;14:251-7.
110. Patrick DL, Erickson P. *Health status and health policy: quality of life in health care evaluation and resource allocation*. New York: Oxford University Press; 1993.
111. Birch S, Gafni A. Cost-effectiveness and cost utility analyses: methods for the non-economic evaluation of healthcare programs and how we can do better. In: Geisler E, Heller O, editors. *Managing technology in healthcare*. Boston, Mass.: Kluwer Academic; 1996.
112. Drummond M, O'Brien B, Stoddart G, Torrance G. *Methods for the economic evaluation of health care programmes*. 2nd ed. Oxford: Oxford University Press; 1997.
113. Ellwood P. Outcomes management: a technology of patient experience. *N Engl J Med* 1988;318:1549-1556.
114. Browne G, Roberts J, Gafni A, Byrne C, Weir R, Majumdar B, et al. Economic evaluations of community-based care: lessons from twelve studies in Ontario. *J Eval Clin Pract* 1999;5:367-85.
115. Nixon J, Stoykova B, Christie J, Glanville JM, Drummond MF, Kleijnen J. The UK NHS Economic Evaluation Database: Economic Issues in Evaluations of Health Technology. *Int J Health Tech Assess* 2000;16:1-12.
116. Harden A, Weston R, Oakley A. *A review of the effectiveness and appropriateness of peer-delivered health promotion for young people*. London: EPPI-Centre, Social Science Research Unit; 1999.