METHODS FOR META-ANALYSIS IN THE PHARMACEUTICAL INDUSTRY

A RECORDED COURSE FOR PSI

Peter Lane

Statistical Consultant
Formerly in GSK's Research Statistics Unit
and Statistical Consulting Group

Developed from courses created and presented within GSK by Peter Lane
with Nick Galwey and George Quartey

September 2015

1. INTRODUCTION AND OVERVIEW

1.1. Emergence of meta-analysis

Ideas behind meta-analysis predate Glass's work by several decades (see, for example, the entry in Wikipedia).

Pearson (1904)
Method for summarizing correlation coefficients from studies of typhoid vaccination.

Tippet (1931) and Fisher (1932)
Presented methods for combining *p*-values.

Yates & Cochran (1938)
Considered the combination of estimates from different agricultural experiments.

R. A. Fisher (1944)
"When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance" (p. 99). Fisher is the source of the idea of cumulating probability values.

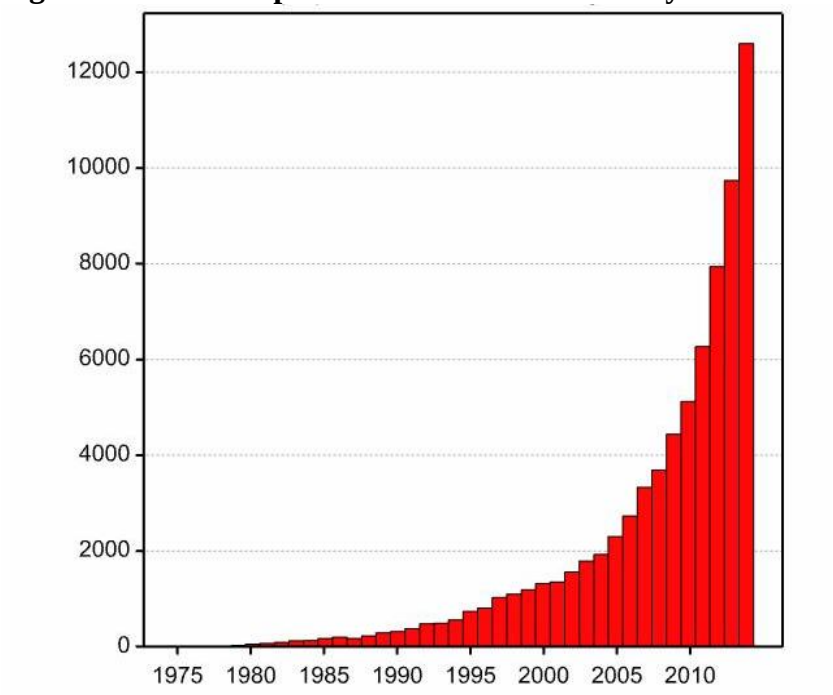
Hans J. Eysenck (1952)
Concluded that there were no favourable effects of psychotherapy, starting a raging debate. Twenty years of evaluation research and hundreds of studies failed to resolve the debate.

W. G. Cochran (1953)
Discussed a method of averaging means across independent studies. He laid out much of the statistical foundation that modern meta-analysis is built upon (e.g., inverse-variance weighting and homogeneity testing).

G.V. Glass (1976)
To prove Eysenck wrong, Gene V. Glass statistically aggregated the findings of 375 psychotherapy outcome studies. He (and colleague Smith) concluded that psychotherapy did indeed work. Glass called his method "*meta-analysis*".

The frequency of meta-analysis publications has increased greatly in recent years (Fig. 1.1, compiled from Google Scholar, June 2015).

Fig. 1.1. Number of publications with 'meta-analysis' in the title, 1976-2014.



1.2. What is meta-analysis?

Meta-analysis combines the results from two or more studies. If used appropriately, it is a powerful tool to summarize results from multiple studies, provides insights into heterogeneous studies, and assists in deriving meaningful conclusions.

Reasons for performing a meta-analysis

- Improving the power to detect a small difference if the individual studies are small
- Improving the precision of the effect measure
- Comparing the efficacy of multiple drugs within a drug class or evaluating the consistency and differences in effect measures across study characteristics
- Evaluating whether overall positive results are also seen in pre-specified subgroups of patients
- Evaluating safety in a subgroup of patients, or a rare adverse event in all patients or when small effects are being assessed when randomized clinical trials (RCT) may be prohibitive; meta-analysis may be used for answering a key safety issue, based on RCTs conducted for efficacy

- Allows investigators to ascertain what data are needed to answer important questions, how many patients should be recruited, and even whether a new study is unnecessary because the questions have already been answered
- Generation of hypotheses for future studies

The role of meta-analysis in the pharmaceutical industry

- Cumulative MA may provide evidence on treatment strategy
- Valuable input into designing a clinical development plan
- Review of a treatment or therapeutic area:
    - How has this class of drugs performed in terms of efficacy and/or quality of life?
    - What can we realistically expect from this drug?
    - What is the safety profile of this class of drugs?
- Planning marketing strategy (in comparative effectiveness research, to support the case for reimbursement)
- Validation of surrogate markers
- Margin selection in non-inferiority studies

PSI's meta-analysis activities

- PSI has a Special Interest Group (SIG) studying Health Technology Assessment, which includes meta-analysis. For example it published an introduction to network meta-analysis (Jones et al 2011).
- The annual PSI conference often includes sessions on meta-analysis, for example in 2010 and 2011.
- An expert group was formed in 2009 to compare meta-analysis standards between the industry and academia, and published a report (Lane et al. 2013) and details of a tool for assessing the quality of meta-analysis (Higgins et al. 2013).

1.3. Common criticisms of MA (and responses)

Meta-analysis ignores qualitative differences between studies.
Meta-analysis does not ignore these differences, but can code them as moderating variables. That way their influence can be empirically tested.

Meta-analysis is a garbage-in, garbage-out procedure.
This is true. However, since the specific content of meta-analyses is always presented, it should be easier to detect poor meta-analyses than it would be to detect poor narrative reviews.

Meta-analysis ignores study quality.
The effect of study quality is typically coded as a moderator, so we can see if there is any difference between good and bad studies. If a difference does exist, low quality studies can be removed from analysis.

Meta-analysis adds together apples and oranges.
The purpose of a literature review is to generalize over the differences in primary research. Over-generalization can occur just as easily in narrative literature reviews as it can in meta-analysis.

Meta-analysis cannot draw valid conclusions because only significant findings are published.
Meta-analyses are actually less affected by this bias than narrative reviews, since a good meta-analysis actively seeks unpublished findings. Narrative reviews are rarely based on an exhaustive search of the literature.

Meta-analysis only deals with main effects.
The effects of interactions are examined through moderator analyses.

Meta-analysis is regarded as objective by its proponents but really is subjective.
Meta-analysis relies on shared subjectivity rather than objectivity. While every meta-analysis requires certain subjective decisions, these are always stated explicitly so that they are open to criticism.

Meta-analysis over-emphasizes small effects.
This is true. Power of detecting effect is magnified by pooling many trials. However, interpretation should give context of clinical relevance.

1.4. Limitations and pitfalls

Some potential disadvantages of meta-analysis arise from combining dissimilar studies or unrepresentative studies. Other pitfalls arise from an incorrect choice of a statistical model and from biases that can be introduced by statistical procedures for pooling, particularly when data are unbalanced between treatments. The following is a list of common pitfalls:

Selection bias

- Ideally, the decision to include a study should be made by looking at its methods and not the results, or by looking at the two separately under blinded conditions
- Meta-analysis should list all studies considered, all studies excluded, and the reasons for exclusion

Publication bias

- Obtaining all published studies is difficult
- Relying on only published studies may distort the results (How hard did they search for all available studies? More than one bibliographic index? Any attempts to consider unpublished data? (E.g. dissertations, conference proceedings, personal contact)

Data-extraction bias

- Interpretation is often required (e.g. reports may list a variety of endpoints, subgroups)
- Ideally, the data should be extracted by more than one observer, each of whom is blinded to various treatment groups.
- Level of agreement between observers should be assessed
- Discrepancies between observers should be resolved

Comparability of studies

- The analysis should:
  - account for the fact that data are from different studies
  - use appropriate model
  - discuss differences or similarities between the individual studies at baseline
  - discuss differences or similarities between the individual study results

Retrospective research

- Changes in diagnostic criteria due to improved technology
- Changes in how the condition is treated
- Changes in concomitant medication
- Quality and availability of original data
- Meta-analysis may be initiated because a signal has already been seen, or guessed at

1.5. When to combine individual studies

Sources of variability in meta-analysis

Decisions to combine studies should be based on thorough investigation of variability among studies, which can be categorized into three types (Higgins and Thompson 2002):

- Variability in study population characteristics, interventions and outcomes is considered *clinical diversity*.
- Variability in study design and quality, such as blinding and concealment of allocation, is considered *methodological diversity*.
- Variability in the observed treatment effects being evaluated in different trials is considered *statistical heterogeneity.*

Statistical tests of heterogeneity help analysts to identify variation among effect estimates (some further details in Section 6). However, basing the decision to combine or not on statistical tests alone is ill-advised. Cochran's Q (see Section 6.1) is the standard test for statistical heterogeneity among studies. This test has low power to detect heterogeneity when the number of studies is relatively low or when individual studies are small; it is sensitive for detecting unimportant heterogeneity when the number of studies is high (Hardy & Thompson 1998). Because of its low power, a value of 0.10 instead of 0.05 is routinely used to determine statistical significance (Higgins & Thompson 2002).

Fig. 1.2 shows a forest plot of 13 trials comparing two treatments for colorectal cancer (Whitehead, 2003). The estimates of treatment effect are fairly evenly scattered, and the variation among them is not large compared with the precision of the individual estimates, so it seems reasonable to combine them. However, the situation is not so straightforward in Fig. 1.3 (a hypothetical example). Each study individually shows a statistically significant difference in favour of the new drug, as illustrated by 95% CIs, all of which lie entirely above 0. Studies 1 and 3 have similar size of effect, but in study 2 the effect is much larger. Consider a meta-analysis of these studies using the methods for combining study estimates described in Sections 5 and 6. The 95% CI based on a fixed-effects model (0.85, 1.33) lies between the two extremes but is not consistent with either: it does not seem an appropriate summary of the results. The test for heterogeneity using the *Q* statistic is highly significant ($Q=70.4$; $p < 0.001$), confirming this concern. But a random-effects analysis indicates that the treatment effect is not significantly different from 0 at the 5% level, although it is significant in every individual trial. Likewise, the 95% CI based on the random-effects model (-0.09, 2.81) is much wider and includes small negative values. Clearly it would be desirable to investigate the studies further, in particular to investigate why the effect in study 2 is so different from the other two. Even though all three studies show a statistically significant benefit for the new drug, there should be concern about the variation in the size of the effect.

Figure 1.2. Forest plot of 13 trials comparing two treatments

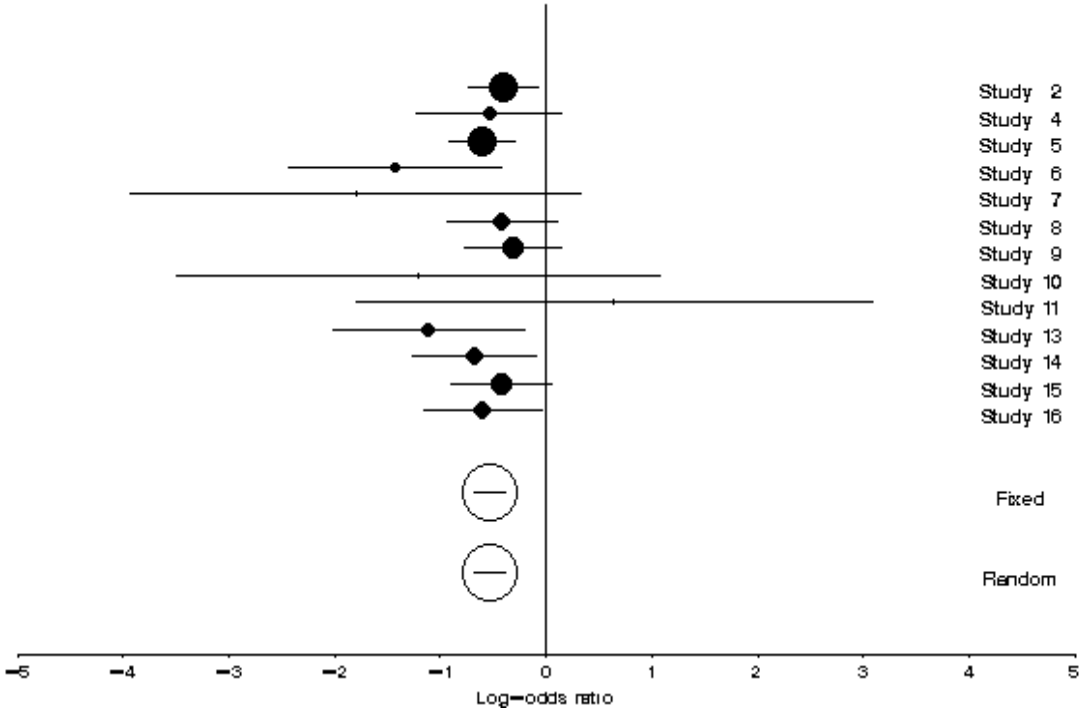
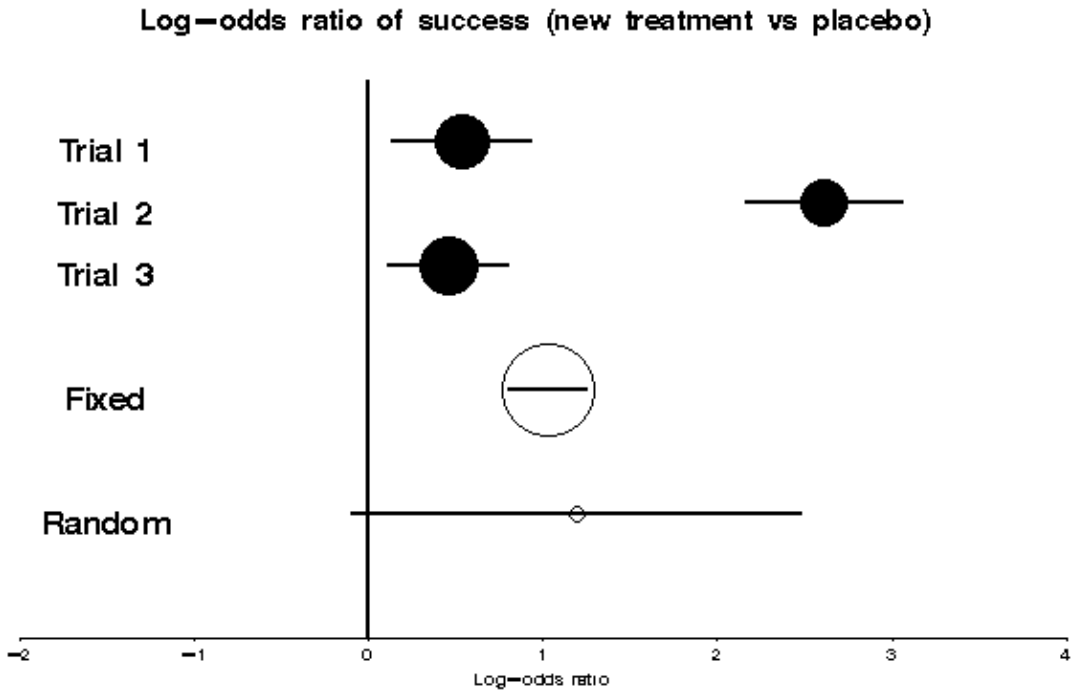


Figure 1.3. Hypothetical example of a meta-analysis.



In deciding whether to combine studies, the most important considerations are whether the studies asked similar questions and whether the study populations are similar enough to yield a meaningful result when they are combined. Unfortunately, no commonly accepted standard exists for "similar enough". Judgment of the similarity among studies depends on the scope of the research question. A more general question may allow more variation among studies than a more focused question. For example, it sometimes makes sense to combine studies from a class of drugs instead of a particular drug – if the drug class in general is of interest, where the included studies were conducted in similar populations in a similar manner, and the drugs in the class affect the outcome in question through similar mechanisms.

Approaches to some common situations are described below.

Combining a small number of studies

There is no general rule for deciding the minimum number of studies for a meta-analysis. The main issue lies in the interpretation of the results. Few will argue with the reliability of a meta-analysis of two mega-trials (5,000 patients or more). However, a meta-analysis of two RCTs with a total of 40 patients should be approached with caution. Even if the combined estimate is statistically significant, the confidence interval is likely to be very wide and could change dramatically with the addition of more studies (also it is unlikely to be judged 'representative' of wide range of conditions or populations). Thus, the results of meta-analyses of a small number of studies should be interpreted cautiously or meta-analysis should be deferred until more studies are available.

Combining studies with different comparators

Even when populations and the intervention under study are homogeneous, comparators may not be. For example, trials might use a "usual care" comparator, a specified comparison therapy, or a placebo group comparator. Not only are these distinct kinds of control groups, but "usual care" may differ across settings and countries or over time. The assumption that all comparators are similar enough to combine needs to be carefully considered. In another situation, a co-intervention is added in all comparison arms of some studies but no co-interventions (or different co-interventions) are added in other studies. For example, anticoagulation might be added to both arms in a trial that compares drug-eluting stents with bare-metal stents. Such a study might be included in a meta-analysis on the effect of anticoagulation, but would provide no within-study information. More generally, the group of trials comparing A vs. B could be described as one group of studies of A + X vs. B + X; others with A+Y vs. B+Y, and so on.

Summarizing studies with different comparisons makes the implicit assumption that no interactions occur between the common added components X or Y and any of the interventions of interest. This assumption needs to be evaluated before meta-analysis is carried out. This type of interaction applies to evaluation of harms as well. A special, common case is when the comparators are different drugs in the same class. Analysts need to consider both the similarity of the comparator drugs and their dosing before

deciding to pool different trials.

The decision to pool data across indications or across classes of drugs requires assumptions which may not be valid. Although it may be true that expanding the range of studies included in an analysis could increase power by including more subjects, it has the potential for masking a signal coming from one drug or class of drugs, or specific to one or more indications, by including studies where there is no signal. Ironically, in an attempt to improve signal detection, combining studies inappropriately may reduce our ability to detect a signal. The reverse effect can also occur: the effect of pooling over potentially heterogeneous populations can be to tar them all with the same brush, when some populations (or classes of drug) give a signal and others don't.

Detailed discussion of the strengths and weaknesses of the assumptions is therefore of critical importance. While it may be reasonable to assume that all members of a pharmacological class, which share the same mechanism of action and are used in the same indication and patient population, have a uniform treatment effect, this is less likely when drugs are pooled across multiple indications or across very different mechanisms of action. Even if one can argue that the direction of treatment effect might be similar, it is a further leap to assume that the effect sizes would be similar. One way to view this problem is to consider some hierarchy of options for combining studies, as follows:

- The ideal approach, if sufficient information is available, is to evaluate data from a single compound in a single indication.
- Next best would be to combine data from similar (mechanism, chemistry) compounds in the same patient population (indication).
- If necessary, it may be useful to combine data from similar compounds across some patient populations
- Only under extraordinary conditions should data be combined for compounds with dissimilar mechanisms of action and from dissimilar patient populations

With a binary response where the data are rich with events, it is easy to define more narrowly which studies will be pooled. It is when events are rare that we need to consider carefully how to expand the scope of our dataset in a responsible manner. When dealing with rare events, it is difficult to detect heterogeneity statistically, so we may have to rely on some concept of clinical heterogeneity. By collapsing across drugs with various mechanisms of action as well as studies of numerous indications, we will run the risk of looking at the effect of treating the disease(s) instead of evaluating the effect of a drug or class of drug.

One of the concerns with combining studies of compounds with dissimilar mechanisms of action or from dissimilar patient populations is that they may have different underlying rates of occurrence of the event of interest. We may want to consider exploring an expanded dataset (i.e. one with data from compounds with dissimilar mechanisms of action or from dissimilar patient populations) using study-level covariates which could allow estimation of the underlying rates as well as the combined risk difference; this is

meta-regression (which we do not cover in this course). This would allow us to look at the overall signal but also evaluate whether differences exist between drugs or classes of drugs. Finally, there is the problem of interpretation: if a meta-analysis produces a pooled estimate over a heterogeneous population, that estimate has less value to inform an individual in one of the populations because any effect relevant to them is diluted or inflated by effects relevant to people in different populations.

Combining studies that use composite outcomes
Composite outcomes need to be viewed carefully in meta-analysis. Composite outcomes bring together two or more events to be considered as a single outcome. The events could be from the same domain – for instance, cardiovascular events such as cardiovascular mortality, non-fatal myocardial infarction, and revascularization. They also can be from different domains with a common cause – for example, a composite endpoint of adverse drug events may include gastrointestinal effects and headache. Finally, they may reflect a common endpoint caused by competing factors – for example, all-cause mortality following coronary artery bypass includes peri-operative deaths (which would be avoided if surgery were not performed) as well late cardiac deaths (which may be reduced by surgery).

In a meta-analysis, one should consider only composite outcomes that are generally agreed upon and in wide usage by the primary studies. Here, creating *de novo* composite outcomes should be avoided.

A composite outcome has the advantage of better statistical power, but it has to make clinical sense. Analysts evaluating the appropriateness of using a composite outcome must take the research question into consideration. A composite outcome with events from the same domain may be justifiable in certain cases, as when included studies reported rare but related adverse events (e.g. stroke, coronary heart disease, and myocardial infarction, which are all in the cardiovascular domain). By contrast, a composite outcome with events from different domains is generally avoided.

1.6. The Cochrane Collaboration

The Cochrane Collaboration is an international network of individuals, launched in 1993, to prepare, maintain, disseminate systematic reviews of research on the effects of health care, promoting Evidence Based Medicine. Their research is made available electronically via the NHS Centre for Reviews and Dissemination (see www.york.ac.uk/inst/crd/). They produce software (RevMan) and journals (Clinical Trials and Meta-Analysis, Evidence Based Medicine).

The Cochrane Library, maintained by the Cochrane Collaboration, produces:

- the Cochrane Database of Systematic Reviews (CDSR, http://www.cochranelibrary.com/cochrane-database-of-systematic-

reviews/index.html)

- Database of Abstracts of Reviews of Effectiveness (DARE, http://community.cochrane.org/editorial-and-publishing-policy-resource/database-abstracts-reviews-effects-dare)

- the Centra Register of Controlled Trials (CENTRAL, http://community.cochrane.org/editorial-and-publishing-policy-resource/cochrane-central-register-controlled-trials-central)

- many other sources of information

The Cochrane approach to meta-analysis of randomized controlled trials stipulates:

- Careful consideration of susceptibility of studies to bias
- meta-analysis when justified rather than by default
- avoidance of fixed-effects model in the presence of heterogeneity
- pre-specification of important potential effect modifiers
- careful interpretation of findings (CIs, not *p*-values)

References

Hardy RJ, Thompson SG (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* **17**:841–856.

Higgins JPT, Lane, PW, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, Haughie S, Hollis S, Lewis SC, Moneuse P, Whitehead A (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods* **4**:351–366.

Higgins JPT, Thompson SG (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**:1539–1558.

Jones B, Roger J, Lane PW, Lawton A, Fletcher C, Cappelleri JC, Tate H, Moneuse P (2011) Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* **10**:523–531.

Lane PW, Higgins JPT, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, Haughie S, Hollis S, Lewis SC, Moneuse P, Whitehead A (2013). Methodological quality of meta-analyses: matched-pairs comparison over time and between industry-sponsored and academic-sponsored reports. *Research Synthesis Methods* **4**:342–350.

Whitehead A (2002). *Meta-analysis of Controlled Clinical Trials*. Chichester: Wiley.
of Internal Medicine **107**:224–233.

Additional reading

Chan A, Altman D (2005). Identifying outcome reporting bias in randomized trials on PubMed: review of publications and survey of authors. *British Medical Journal* **330**:753.

Chan A, Hrobjartsson A, et al. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *JAMA* **291**:2457–2465.

Chan A, Krleza-Jerić K, et al. (2005). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* **171**:735–740.

Cochran handbook (accessed June 2008), URL http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.doc

Copas J, Jackson D (2004). A bound for publication bias based on the fraction of unpublished studies. *Biometrics* **60**:146–153.

Gotzsche PC (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of non-steroidal anti inflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials* **10**:31–56.

Hayashi K, Walker AM (1996). Japanese and American reports of randomized trials: differences in the reporting of adverse effects. *Control Clinical Trials* **17**:99–110.

Ioannidis JPA, Lau J (2001). Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* **285**:437–443.

Ioannidis JPA, Lau J (2002). Improving safety reporting from randomized trials. *Drug Safety* **25**:77–84.

Williamson PR, Gamble C (2007). Application and investigation of a bound for outcome reporting bias. *Trials* **8**:9.

2. STEPS IN CONDUCTING META-ANALYSIS

2.1. Study objectives

You need to clearly state, and agree, the objectives of a meta-analysis at the start of planning, as for a clinical trial. A check-list was given by Counsell (1997).

- Population: condition(s), disease severity and stage, co-morbidities, patient demographics
- Intervention (or treatment): dosage, frequency, and method of administration
- Comparator: placebo, usual care, or active control
- Outcome: FEV1, heart attack, mortality, quality of life
- Timing: duration of follow-up
- Setting: primary, specialty, in-patient; co-interventions

Start by understanding the clinical and policy decisions that the analysis is intended to inform. Discussion with stakeholders and content experts can help to define key issues involving details of the intervention, specific sub-populations of interest, and outcomes of importance.

In addition, you need to decide the types of study to be included in the meta-analysis: (observational, RCT, blinded).

2.1.1. Example: individual patient data

This is taken from GSK's analysis plan for meta-analysis of Avandia in 2006 (Cobitz et al 2008).

Primary objective:
Investigate whether a relationship exists between rosiglitazone and congestive heart failure events or myocardial ischaemia events while controlling for baseline characteristics.

The analyses compared rosiglitazone (RSG) to active or placebo control, separately for the following treatment regimens:

1. as monotherapy,
2. in combination with metformin (Met),
3. in combination with a sulphonylurea (SU),
4. in triple therapy (Met+SU+RSG), or
5. in combination with insulin.

The analysis plan also specified that only RCT data with blinded comparison was to be used. This is a frequently used restriction made in meta-analysis, to focus on what is perceived as the highest quality of information; Section 2.3 looks at that in some more

detail. In addition, comparator regimens are divided, there was no mention of duration of trials, and the population was of "diabetic patients".

2.1.2. Example: summary data

From Nissen & Wolski's meta-analysis of Avandia (Nissen & Wolski, 2007)

- "We performed a meta-analysis of trials comparing rosiglitazone with placebo or active comparators to assess the effect of this agent on cardiovascular outcomes."
- "Criteria for inclusion in our meta-analysis included a study duration of more than 24 weeks, the use of a randomized control group not receiving rosiglitazone, …"
- "Table 2 reports the doses of rosiglitazone and comparator drugs …" and also the population targeted in each trial.

Notes

- blinding not required
- population not specified
- all comparator regimens lumped together

2.2. Choice of outcomes

Most meta-analyses are of treatment differences or odds ratios. However, any statistic that summarizes the effect of an intervention can be used in a summary-level analysis, and any outcome can be used in a patient-level analysis. Chapter 3 looks in detail at the more common choices of effect measure. The most important factor in choosing the statistic that it should characterize the effect of the interventions to be compared in a way that is likely to be reasonably consistent across studies.

In summary-level analysis, you need the chosen statistic for each contributory trial. If a trial report does not provide it, then that trial can only be included in the analysis if you can get access to the data in order to calculate it. When working with binary outcomes, however, the information about how many events there were in each treatment arm for how many patients is sufficient to calculate simple alternative statistics, such as the risk difference or the odds ratio.

In addition, you need the standard error of the statistic. If this is not provided explicitly, it is possible to derive approximate values using established formulae relating standard errors to other measures that may be reported, such as ranges or quartiles (as used in boxplots, for example). This will depend on the distribution of the statistic: it is best to work with a statistic that has an approximately Normal distribution. So standard methods of meta-analysis of odds ratios and relative risks, for example, are applied to the log-odds and log-relative risks; the usual methods of estimating these from individual trials produces estimates on the log scale with SEs on that scale as well. It is particularly

important combining statistics whose reciprocals have an approximately Normal distribution, as their behaviour can be difficult to characterize. An example is the NNT, or number need to treat, which is a reciprocal of the risk difference for a binary outcome. NNT should not be combined directly in a meta-analysis: instead, the risk differences can be combined (if that is likely to be consistent across studies) and the resulting combined risk difference expressed as an NNT for the purposes of interpretation. If the odds ratios are more likely to be consistent across studies, they can be combined, and the combined odds ratio expressed as a risk difference or NNT.

If trials report different outcomes, you need to take care about combining them. It is possible to use effect sizes to handle this problem. However, this only makes sense if you can justify from a clinical point of view that the different outcomes can reasonably be held to measure a common aspect of the disease, whose summary will be informative about the studied interventions.

Meta-analysis is sometimes carried out on simple means or proportions, rather than on comparative measures like differences or odds ratios. This may be all that is possible, for example to compare to historical data. However, it is essential to recognize that this approach loses one of the main benefits of standard methods of meta-analysis: the protection of the randomization structure of the individual trials. When absolute statistics like means are compared across trials, there is always likely to be doubt whether any differences found are due to treatment differences or to differences between the trials.

2.3. Type of study to include

The main issue with study selection is that the inclusion of studies that produce unreliable estimates may lead to biased results. Various approaches have been taken to systematize the assessment of quality of studies: see Sutton et al, Chapter 8 for details. They include the following hierarchy of evidence, based on the methodology behind the study.

1. Well-designed randomized controlled trial

2. Other types of trial
   - 2.1 Well-designed controlled trial without full randomization
     - 2.1a Well-designed controlled trial with pseudo-randomization
     - 2.1b Well-designed controlled trial with no randomization
   - 2.2 Cohort studies
     - 2.2a Well-designed cohort (prospective) study with concurrent controls
     - 2.2b Well-designed cohort (prospective) study with historical controls
     - 2.2c Well-designed cohort (retrospective) study with historical controls
   - 2.3 Well-designed case-control (retrospective) study

3. Large differences from comparisons between times and/or places with and without intervention (in some circumstances, these may be equivalent to Level 2 or 1)

4. Opinions of respected authorities based on clinical experience; descriptive studies, and reports of expert committees

2.3.1. Types of evidence

Excerpt from the EPC Manual for Conducting a Systematic Review (1996) (Woolf 1996).

> Collecting and reading the literature is one of the most time-consuming tasks in a systematic review. Expending these resources can be especially wasteful if the reviewers "cast too wide a net" and gather evidence of poor quality or with limited relevance to the questions raised by the evidence model. On the other hand, if the literature review is too narrow, important sources of evidence may be omitted (Slavin, 1995).
>
> Published evidence can include a heterogeneous group of data sources of variable quality and relevance. Excluding an entire category of literature is not without risks. Randomized controlled trials are unavailable for many aspects of medicine, due largely to the cost and time requirements to perform them. Limiting a review to such trials might exclude important data from other types of studies (e.g., cohort studies, case-control studies, descriptive epidemiology).
>
> For some topics, evidence from animal models or laboratory studies is essential. Even review articles, editorials, and letters-to-the-editor, which are often omitted because they lack primary research data, can provide important insights about published studies. Their reference lists can also help verify the comprehensiveness of the review's bibliographic database.

On the other hand, casting a wide net opens the door to studies of dubious quality and can expand the volume of a search to hundreds or thousands of superfluous articles. Doing so can be especially inefficient. If good evidence from a few major clinical trials is available, there may be no purpose in spending time and money to collect hundreds of retrospective studies and case reports on the same subject. Thus, before they determine the appropriate boundaries for admissible evidence, reviewers should conduct a preliminary literature search to obtain a sense of the type of evidence that is available. They can then perform a "best-evidence" synthesis, limiting the review to the highest quality studies and foregoing the collection and review of other evidence (Slavin, 1995).

2.3.2. Clinical trials

Studies may use cross-over as well as parallel designs. This can be difficult to handle for patient-level data, though straightforward for the analysis of summary data. However, when summarizing cross-over studies, it is important to use standard errors of differences

from within-patient variability rather than from between-patient.

When studies have multiple treatment arms, the simplest approach is to analyse treatments in pairs. A more powerful approach is to include all treatments, to benefit from indirect comparison (Section 7). With multiple doses:

- Compare pairs of dose-levels as for treatments
- Common dose-levels: analyse as for multiple treatments (benefit from indirect comparison)
- Varying dose-levels: need to model dose-response as part of analysis
- Titrated treatments make a formal quantitative approach hard

Many clinical trials are carried out at multiple centres. When combining patient-level data from such trials, many strategies are possible. It is important to recognize that from a statistical point of view analysis of multi-centre trials is analogous to meta-analysis anyway (Senn, 2000). If the effect of centres is important, one approach is to combine estimates from each centre in each trial, so that "centre" becomes "study".

Sometimes, clinical trials have a factorial design, intended to estimate the effect of more than one factor in a single trials; for example, this is particularly relevant when studying the effect of doses of combinations of drugs. For a patient-level analysis, the factorial model needs to be extended to include studies, as with studies with multiple treatment arms. If the designs of the studies are different, a hierarchical approach is possible, but difficult (Frost et al, 1999). For a summary-level analysis, standard estimates of the effects of interest, whether of main effects, interactions, or specified contrasts, can be combined as for simpler designs.

The combination of results from studies with adaptive designs is complicated by alpha-adjustment due to interim analyses. One solution is to combine p-values (Fisher, 1932).

2.4. Search strategy

For a prospective MA, the identification of studies to combine is no problem, and for an in-house MA it should be straightforward. The main issue is combining published material. The search can be expected to be time-consuming, and require a comprehensive strategy using several sources. Clinical trial registers are helping this process, and there are central registers in the USA that are required by regulation. One of the main problems is the possibility of publication bias. The sources and methods used in searching, however, are too extensive and detailed to cover in this course. An excellent resource on this subject, however, is the Cochran handbook, Chapter 5.

2.5. Combining the information

For summary data, the commonest method of calculating a combined estimate in a meta-

analysis is the "inverse-variance" method. All that is required is an estimate of some statistic from each study, such as a treatment difference or an odds ratio, together with its standard error. The method uses the inverse variance (i.e. $1/SE^2$) to weight each estimate. So studies with imprecise estimates (i.e., with large SEs) will make less contribution to the combined estimate than those with precise estimates.

One of the main components of the SE of the estimate from a study is the number of patients involved. In general, the SE decreases as the number increases, but is inversely proportional to the square root of the number. Hence, using the variance rather than the SE for the weighting means that studies contribute more or less in proportion to the number of patients in the trial. Of course, the SEs will also depend on the individual variability in each trial, which will itself vary from trial to trial because of other factors, such as the population of subjects and the operating procedures of the trial.

Together with the combined estimate itself, the inverse-variance method provides a standard error of that estimate, and so confidence intervals and a p-value can be derived. In addition, the heterogeneity of the estimates can be calculated, relative to the variability of the estimates, to provide information about how consistent the individual contributions are. This heterogeneity is central in the interpretation of the meta-analysis, and can be used to help decide whether the studies should all be combined in the first place. This will be discussed further in Section 7.

To illustrate the process, here is a meta-analysis of nine RCTs comparing two dentifrices (toothpastes or tooth powders), containing NaF (sodium fluoride) or SMFP (sodium monofluorophosphate). (Taken from Sutton et al, 2000, Page 30, and originally in Johnson MF, 1993.)

Table 2.1. Outcome of nine trials comparing two dentifrices.

Study	N1	Mean1	SD1	N2	Mean2	SD2
1	134	5.96	4.24	113	6.82	4.72
2	175	4.74	4.64	151	5.07	5.38
3	137	2.04	2.59	140	2.51	3.22
4	184	2.70	2.32	179	3.20	2.46
5	174	6.09	4.86	169	5.81	5.14
6	754	4.72	5.33	736	4.76	5.29
7	209	10.10	8.10	209	10.90	7.90
8	1151	2.82	3.05	1122	3.01	3.32
9	679	3.88	4.85	673	4.37	5.37

Here, Treatment 1 is NaF and 2 is SMFP, and the means and SDs are of the outcome variable which is difference from baseline in the DMFS score (Decayed, Missing and Filled Surfaces in secondary teeth). The SDs in the two treatment arms in each study, s_1 and s_2, are similar, as is usually expected to be the case, and can be pooled using a simple formula:

$$\text{Pooled SD} = \sqrt{\{ [(n_1-1)\, s_1^2 + (n_2-1)\, s_2^2]/(n_1+n_2-2)\}}$$

The treatment differences, d_i (i=1…9), and SEDs can then be calculated, using

$$\text{SED} = \text{SD}*\sqrt{\{1/n_1 + 1/n_2\}}$$

and the weights, w_i, for combination of the estimates are the inverses of the squared SEDs.

Table 2.2. Treatment differences in the dentifrice studies.

Study	Diff	SED	Weight
1	0.86	0.57	3.07
2	0.33	0.55	3.25
3	0.47	0.35	8.09
4	0.50	0.25	15.88
5	–0.28	0.54	3.43
6	0.04	0.28	13.21
7	0.80	0.78	1.63
8	0.19	0.13	55.97
9	0.49	0.28	12.92

Notice that Study 8 has the highest weight – as expected from the number of subjects. In fact, it has nearly as much weight as all the other trials put together, so the combined estimate will depend heavily on the estimate from that one study. By contrast, Study 7 is hardly going to contribute to the combined estimate at all with a weight of less than 2 – this is mostly because of the high level of variability in that study, together with the relatively small number of patients.

So the combined estimate is $\sum w_i d_i / \sum w_i = 0.28$, and its SE is $\sqrt{\{1/\sum w_i\}} = 0.09$. Using the usual multiplier of 1.96 to calculate a 95% confidence interval for a Normally distributed estimate gives (0.10, 0.46) and a p-value of 0.002.

This shows that even though only one of the nine studies showed a significant difference between treatments (Study 4), the combined evidence is significant – as long as the studies are suitable for combination (as discussed earlier). Again subject to the caveats of combination (see Section 1), an estimate of the average superiority of SMFP over NaF is an additional decrease from baseline of 0.3 units on the DMFS scale.

For patient-level data, the methods of combining information over studies are closely analogous to the methods used in analysing a single multi-centre trial. All the data can be handled together in a single analysis, rather than analysing each trial separately and then combining. The difficult part is to decide what model to fit. The main components are straightforward, as the outcome variable and the type of analysis (e.g. survival analysis, logistic regression, mixed model) are likely to be suggested from the original trials, though there may be difficulties if the trials were analysed differently. In that case, there may be no solution except to resort to combining summary information. But the biggest

difficulty may well be the selection of effects to include in the model. Clearly, the effect of Treatment must be fitted, but there is the opportunity to fit additional covariates which may account for some of the heterogeneity between studies. Of course, not all covariates may be recorded in all studies, and this raises another problem to overcome, deciding the pay-off between including all studies and including all covariates of interest. Finally, the effect of Study should also be fitted. If this is not done, and there is any imbalance between numbers of subjects on different treatment arms in some trials, there is a risk that the Treatment effect will be partially confounded with the Study effect. The result can be seriously misleading, as we shall see in the examples of Section 5.

There are many other ways of going about the combination of summary information, apart from the simple method above. For example, we can use a random-effects model rather than a fixed-effects model, to put the results in the context of a population of potential trials, rather than just the trials we have collected. Or we can use Bayesian methods rather than frequentist ones. And of course the details of carrying out the combination are different for the various outcome measures. We will look at all of these later.

2.6. Sensitivity analysis

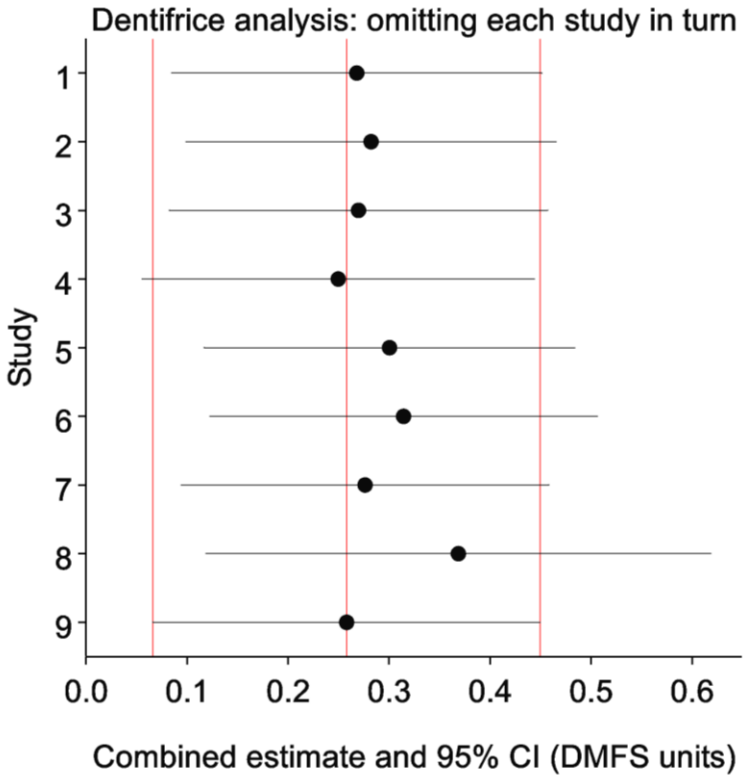
Sensitivity analysis provides information to decide how much the results of the original meta-analysis depend on the assumptions made and decisions taken. The more the results from the meta-analysis are unchanged by sensible modifications to the processes, methods and analyses, the more confident we can be.

Natural candidates for applying sensitivity analysis are as follows.

- Inclusion criteria; omitting some of the chosen studies, or including some that were not originally chosen.
- Adjusting for potential publication bias: how much bias would there have to be to change the conclusions?
- Handling of missing values (see Sutton et al, Chapter 13).
- Alternative methods, including random-effects versus fixed-effects models.

One relatively straightforward check is to run the meta-analysis excluding just one of the chosen studies, for each of the studies in turn, and graph the resulting combined estimates. Once the dataset and programming have been set up for the main analysis, it takes not much more effort to repeat it many times. Figure 2.1 shows the plot corresponding to this approach for the dentifrice analysis, in which the vertical reference lines mark the original estimate and 95% confidence limits.

Figure 2.1. Sensitivity analysis of the dentifrice meta-analysis.

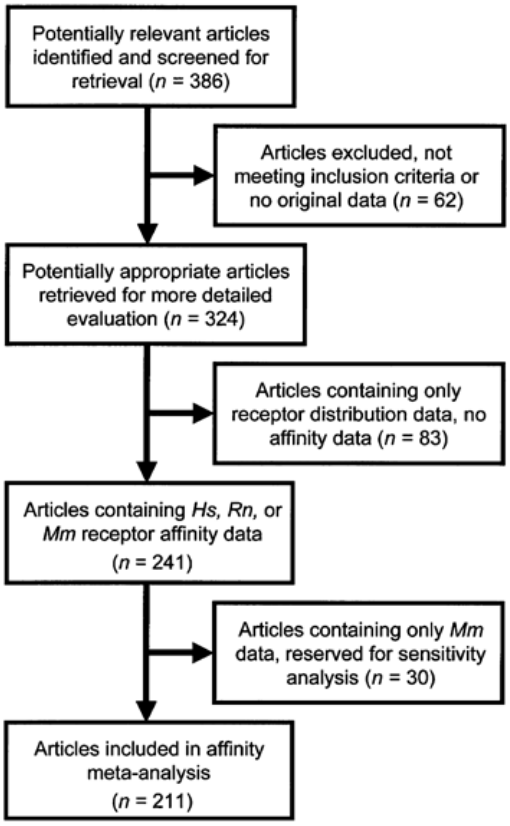


2.7. Interpretation and presentation

Several concerted attempts have been made to improve the standard of reporting of meta-analyses. The Cochrane Collaboration has extensive guides. The CONSORT statement was published (Begg et al, 1996) to improve the standard of reporting of RCTs, and this was followed by the QUORUM statement (Moher et al, 1999) specifically for reporting meta-analyses. The main headings in the checklists included in these two statements are the same. More recently, the QUORUM statement has been updated and expanded to the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which can be found on the CONSORT website, and various publications such as Moher et al (2009). A PSI Expert Group carried a comparison of industry and academic meta-analyses and published a report (Lane et al, 2013) and guidelines (Higgins et al, 2013).

One of the useful items of the Quorum statement is the flow diagram, which records the process of screening potential studies, and the reasons for excluding some of those found. Figure 2.2 shows an example, from McPartland, Glass & Pertwee (2007).

Figure 2.2. QUORUM-style flow diagram.

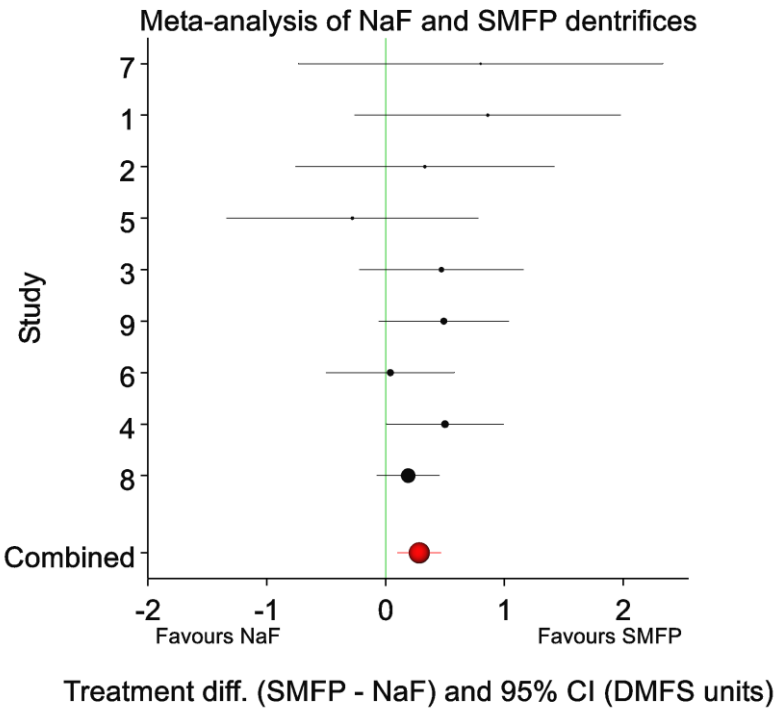


It is widely accepted that any report should include at least two lists of information categorized by study. The first lists the details of the studies themselves, identifying the studies and detailing the treatment and other relevant characteristics of the design such as treatment duration, dosing, numbers of patients, and so on. The second lists the statistical summaries from each study, such as treatment means or event rates – these should be the numbers that are analysed with whatever meta-analysis method is employed. Providing them in this form enables readers to do their own analyses, perhaps with different methods, to assess the results.

The information from this second table is often presented also in graphical form, usually together with the main goal of the analysis: the combined estimate. A graph makes the information much easier to appreciate, and the main features of the component data to be assessed. This form is an interval plot, also called a forest plot, which is a simple dot-plot with error bars. (Its design is similar to that of the sensitivity graph of the last section, with some additions.) In its application to meta-analysis, it is usually referred to as a forest plot (though no-one is clear on the origin of the name, apart from a vague idea that the series of lines might be seen as trees scattered in a forest). Figure 2.3 shows a forest

plot of the dentifrice example.

Figure 2.3. Forest plot of the dentifrice meta-analysis.



We will look at the details of the forest plot, and other graphs that are suitable for reporting meta-analyses, in Section 4.

The type of statistic used to combine information in a meta-analysis should be chosen on statistical grounds. It is important to use a statistic that measures the type of treatment comparison that can be expected to be reasonably similar across the studies, and the statistic needs to have amenable properties for formal analysis. However, this does not mean that this statistic will also be the most appropriate for reporting the results. In fact, it is often the case that a statistic like the odds ratio will be chosen for the calculations, whereas a report in terms of risk difference is far more useful and informative to the end-users of the meta-analysis: the clinicians and patients considering a medical intervention.

It is not difficult to produce an informative summary on a different scale. But first, it is necessary to decide a population, or series of populations, for whom the summary is to be constructed. A meta-analysis may include studies on very different types of patient, and baseline measures of a disease, or risk of some event, are likely to vary widely between studies. The implicit model assumed in the meta-analysis may assume a constant effect of treatment on some scale, but this will not be constant on another. We will look at this in detail in Section 5.

References

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF (1999). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Journal of the American Medical Association* **276**:637–639.

Cochran handbook (accessed June 2008), URL
http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.doc

Counsell C (1997). Formulating questions and locating primary studies for inclusion in systematic reviews. *Annals of Internal Medicine* **127**:380–7.

Fisher RA (1932). *Statistical methods for Research Workers (4th Ed)*. Oliver & Boyd: London.

Frost C, Clarke R, Beacon H (1999). Use of hierarchical models for meta-analysis: experience in the metabolic ward studies of diet and blood cholesterol. *Statistics in Medicine* **18**:1657–1676.

Higgins JPT, Lane PW, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, Haughie S, Hollis S, Lewis SC, Moneuse P, Whitehead A (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods* **4**:351-366.

Johnson MF (1993). Comparative efficacy of NaF and SMFP dentifrices in caries prevention: a meta-analytic overview. *Caries Research* **27**:328–36.

Lane PW, Higgins JPT, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, Haughie S, Hollis S, Lewis SC, Moneuse P, Whitehead A (2013). Methodological quality of meta-analyses: a matched-pairs comparison over time and between industry-sponsored and academic-sponsored reports. *Research Synthesis Methods* **4**:342-350.

McPartland JM, Glass M, Pertwee RG (2007). Meta-analysis of cannabinoid ligand binding affinity and receptor distribution: interspecies differences. *British Journal of Pharmacology* **152:**583–593.

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet* **354**:1896–1900.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed.1000097.

Nissen SE, Wolski K (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine* **356**: 2457–2471.

Senn S (2000). The many modes of meta. *Drug Information Journal* **34**:535–549.

Slavin RE (1995). Best evidence synthesis: an intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology* **48**(1): 9–18.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000). *Methods for Meta-analysis in Medical Research*. Wiley: Chichester.

Woolf SH (1996). *Manual for conducting systematic reviews*. Agency for Health Care Policy and Research, AHRQ: 77pp.

3. CHOICE OF EFFECT MEASURES AND MODEL

3.1. Choice of effect measures for combining studies

Effect measures quantify differences in outcomes between treatments in trials, or between exposure groups in observational studies. Effect measures can be broadly classified into two types:

- *absolute*, e.g., risk difference (RD), rate difference, or mean difference; and
- *relative*, e.g., odds ratio (OR), relative risk (RR), or hazard ratio (HR).

The choice of effect measure in meta-analysis is often prescribed by four factors:

(i) study design and the type of outcome data used, e.g., continuous, dichotomous, ordinal, interval, counts, or time to event,
(ii) the corresponding measure reported, e.g., mean difference or standardized mean difference, relative risk, rate ratio, odds ratio, risk difference, or hazard ratio,
(iii) rarity of the event and
(iv) heterogeneity across baseline risk within a study or heterogeneity in at-risk populations between studies.

3.1.1. Dichotomous outcomes

For dichotomous outcomes, relative effects metrics are generally more appropriate for meta-analyses than the risk difference.

Relative measures (RR, OR) are more likely to be homogeneous across studies, particularly when variation among control group rates is large. Table 3.1 shows how the same RR can correspond to very different values of RD, depending on the control group rate. However, when the control event rates are similar among trials, risk differences may be combined. When rates among control groups vary widely among studies, the investigators may calculate pooled estimates of both measures to see whether pooling the risk differences introduces bias.

Table 3.1. The effect of differences in baseline risk on the risk difference when the odds ratio is constant.

Mild Baseline			
	Event	**No event**	**Total**
Active	5	95	100
Placebo	10	90	100
Total	15	185	200

Severe Baseline			
	Event	**No event**	**Total**
Active	10	90	100
Placebo	20	80	100
Total	30	170	200

Overall			
	Event	**No Event**	**Total**
Active	15	185	200
Placebo	30	170	200
Total	45	355	400

RD = (5/100)-(10/100) = -0.050	RD = (10/100)-(20/100)) = -0.100	RD = -(15/200)-(30/200) = -0.075
RR = (5/100)/(10/100) = 0.500	RR = (10/100)/(20/100) = 0.500	RR = (15/200)/(30/200) = 0.500

Illustrative example

Consider the Stroke data set (Collins et al., 1990) in Table 3.2. Three parameterizations of the treatment difference can be considered, namely the log-odds ratio, the probability difference and the log-relative risk: these are compared in Table 3.3. Fixed-effects meta-analysis based on study estimates can be performed for each parameterization. When the analysis is performed on the basis of the log-odds ratio or the log-relative risk as a measure of treatment difference, the test for heterogeneity was not significant (Table 3.4). However, on the basis of the probability difference there is significant heterogeneity.

Table 3.2. Distribution of events (strokes) in a set of trials to compare a new anti-hypertensive drug with a placebo

Trial	Number of patients				Percent with event	
	Active arm		Control arm			
	Event	Total	Event	Total	Active	Control
1 VA-NHLB1	0	508	0	504	0.0	0.0
2 HDFP (Stratum I)	59	3903	88	3922	1.5	2.2
3 Oslo	0	406	5	379	0.0	1.3
4 ANBPS	13	1721	22	1706	0.7	1.3
5 MRC	60	8700	109	8654	0.7	1.2
6 VAII	5	186	20	194	2.6	9.3
7 USPHS	1	193	6	196	0.5	3.0
8 HDFP (Stratum II)	25	1048	36	1004	2.3	3.5
9 HSCSG	43	233	52	219	15.6	19.2
10 VAI	1	68	3	63	1.4	4.5
11 WOLFF	2	45	1	42	4.3	2.3
12 Barraclough	0	58	0	58	0.0	0.0
13 Carter	10	49	21	48	16.9	30.4
14 HDFP (Stratum III)	18	534	34	529	3.3	6.0
15 EWPHE	32	416	48	424	7.1	10.2
16 Coope	20	419	39	465	4.6	7.7
Total	289	18487	484	18407	1.5	2.6

Table 3.3. Different effect measures in the stroke data.

Study	Percent with event		OR	RR	RD
	Active	Control			
1 VA-NHLB1	0.0	0.0	–	–	0.0000
2 HDFP (Stratum I)	1.5	2.2	0.67	0.68	-0.0071
3 Oslo	0.0	1.3	0.00	0.00	-0.0130
4 ANBPS	0.7	1.3	0.59	0.59	-0.0052
5 MRC	0.7	1.2	0.55	0.55	-0.0056
6 VAII	2.6	9.3	0.26	0.28	-0.0673
7 USPHS	0.5	3.0	0.17	0.17	-0.0245
8 HDFP (Stratum II)	2.3	3.5	0.67	0.67	-0.0113
9 HSCSG	15.6	19.2	0.78	0.81	-0.0361
10 VAI	1.4	4.5	0.31	0.32	-0.0310
11 WOLFF	4.3	2.3	1.87	1.83	0.0193
12 Barraclough	0.0	0.0	–	–	0.0000
13 Carter	16.9	30.4	0.47	0.56	-0.1349
14 HDFP (Stratum III)	3.3	6.0	0.52	0.54	-0.0278
15 EWPHE	7.1	10.2	0.68	0.70	-0.0303
16 Coope	4.6	7.7	0.57	0.59	-0.0318

Table 3.4. Fixed-effects MA of stroke anti-hypertensive data comparing a new drug with placebo, using three effect measures.

Effect Measure	Q-statistic (d.f. = 12)[1]	*p*	Estimate	95% CI		Back-transformed estimate	Back-transformed CI	
				Lower	Upper		Lower	Upper
log(OR)	8.28	0.76	-0.522	-0.672	-0.372	0.593	0.511	0.689
log(RR)	8.90	0.71	-0.487	-0.628	-0.345	0.615	0.534	0.708
RD	22.14	0.036	-0.00684	-0.00920	-0.00448	-0.00684	-0.00920	-0.00448

[1]Significance test for heterogeneity, with a chi-square distribution. d.f. = No. of studies – 1. Studies with zero counts cannot be included in the calculation of this statistic. For further detail see Section 6.2.

In this data set that the percentage of strokes in the control group varies from 0 to 30.4. On the whole the estimates of the log-relative risk and log-odds ratio are similar, but those of probability differences are not. Heterogeneity in the probability difference scale is likely to arise if the control rates take a wide range of values, or if all the rates are close to 0% or close to 100%. For example, if the control rate is 30.4%, a reduction on 0.05 in the probability difference scale leads to a rate in the treated group of 25.4%. If the control rate is 1.3%, the same reduction on the probability difference scale leads to a rate of -3.7%, which is not possible: the largest possible difference in this situation is 0.013. In this example it is perhaps more plausible that the treatment will reduce the rate by multiplicative factor, for example reduce the rate to 90% of the control rate. The log-odds ratio is a more satisfactory measure in this respect.

3.1.2. Continuous outcomes

The choice of effect measure to combine for continuous outcomes is determined primarily by the form of data available. If multiple trials report results using the same or similar scales, mean differences or differences in change between groups can be combined. Alternatively, standardized effect sizes expressed as differences, or differences in change over time, divided by standard deviations are sometimes combined. This method is typically used when outcome measures are reported on different scales (See Section 9.3). This measure can incorporate multiple scales, but results can be difficult to interpret.

3.1.3. Time-to-event outcomes

The effect measure for analysing time-to-event or survival data is the hazard ratio (HR). The HR, an estimate of the relative risk over an infinitesimal time interval, is also referred to as the relative hazard (RH). The most common survival analysis yielding a HR is the Cox proportional hazards model. The model assumes that the HR is constant over time, yielding a single value for a given study. Studies reporting a HR from the model should state explicitly whether or not the proportional hazard assumption was satisfied. However, a HR may not always be available or explicitly reported. Commonly, event rates are reported at various times during follow-up. Under such circumstances, the HR and its variance can be calculated if observed and expected events can be extracted (Parmar, Torri et al., 1998).

3.1.4. Number needed to treat or harm

The number needed to treat (NNT) is an estimate of the number of individuals who must receive the active treatment in order to produce one more response. NTT, and the corresponding number needed to harm (NNH) in the case of adverse events, which are the inverse of the risk difference, may also be considered an effect measure. Estimating and interpreting the NNT or NNH in a meta-analysis may not be straightforward (Quartey *et al*, 2007; Altman and Deeks, 2002; Cates, 2002).

NNTs and NNHs are frequently used because they portray the absolute effect of an intervention in an intuitive way. NNTs and NNHs themselves do not reflect variations attributable to underlying event rates; and they do not have a standardized unit of time.

These drawbacks should be considered when NNTs or NNHs are presented. Report these measures with an appropriate time frame and confidence intervals and make clear that they are based on an average estimate, for example, "On average, 10 patients would have to be treated for 3 years with treatment A to observe one fewer event after 3 years" (Hutton 2000). If substantial variations in NNTs (NNHs) exist based on different event rates, dosages, or subgroups, then these measures should be reported separately for each group.

3.2. Choice of model for combining studies

Meta-analysis can be performed using either *a fixed-effect* or *a random-effects* model. Either type of model can be used to combine effect measures for any type of data, whether continuous, binary, time-to-event or other.

The *Fixed-effect (FE)* and *Random-effects (RE)* models represent two conceptually different approaches. Under the FE model we assume that there is one *true effect size* which is shared by all the studies included. Put another way, all factors which could influence the effect size are the same in all the study populations, and therefore the effect size is the same in all the study populations. It follows that the observed effect size varies from one study to the next only because of the *random error* contributing to individual observations that is inherent in each study.

By contrast, under the *RE* model we allow that the true effect could vary from study to study. For example, the effect size might be a little higher if the subjects are older, or more educated, or healthier; or if the study used a slightly more intensive or longer variant of the intervention; or if the effect was measured more reliably; and so on. The studies included in the meta-analysis are assumed to be a random sample of the relevant distribution of effects, and the combined effect estimates the mean of this distribution.

3.2.1. Which model should we use?

The selection of a model should be based on the nature of the studies and objectives and should not be solely based on tests of heterogeneity. Additional criteria such as the number of trials and the distribution of the individual study estimates of treatment difference need to be considered.

For a meta-analysis based on small number of studies, the estimate of heterogeneity from data is likely to be unreliable. If the results of trials appear to be reasonably consistent the FE analysis may be the more appropriate one to present. If there is inconsistency then no overall estimate should be calculated and a further investigation into the cause of inconsistency needs to be undertaken (see Section 7). For MA based on a larger number of trials the RE analysis may be preferred (see Section 6). However, if the distribution of trial estimates is far from the assumed Normal distribution then further investigation needs to be undertaken.

The overall estimate from FE analysis provides a summary of results obtained from a particular sample of patients contributing the data. A common argument in favour of RE model is that it produces results that are generalizable – though only as far as the population represented by the studies collectively. An important feature of RE is that it allows the between-study variability to influence the overall estimate and more particularly its precision. That is, if there is substantial variability among studies this will

be reflected in the CI of the estimate. Therefore a treatment effect that is significant in the FE analysis may be non-significant in the RE analysis: this is the price that is paid for the generalizability of the RE results. If the variability among trials is substantial, a larger number of trials will be required to demonstrate significance: the same number of trials with a larger number of patients will not help much.

In many cases it is useful to consider the results of both FE and RE model. If there is no heterogeneity, then the RE and FE analysis will be the same because τ^2 (the variance component due to heterogeneity) will be estimated to be 0. On the other hand, if the two analyses lead to important differences in conclusion, this highlights the need for further investigation.

The FE model makes sense if there is reason to believe that all the studies are functionally identical, and the goal is to compute the common effect size, which would then be generalized to other individuals in this same population. For example, assume that a drug company has run five studies to assess the effect of a drug. All studies recruited patients in the same way, used the same researchers, dose, and so on, so all are expected to have the identical effect (as though this were one large study, conducted with a series of cohorts). Also, the regulatory agency wants to see if the drug works in this one population. In this example, a fixed-effects model makes sense.

A FE model may be used to summarize all the available evidence as far as it goes, and this is a valid approach whether or not there is heterogeneity. But such a summary is not useful for assessment of future use of the treatment if there is substantial variation among studies.

By contrast, when the researcher is accumulating data from a series of studies that had been performed by other people, it would be unlikely that all the studies were functionally equivalent. Typically, the subjects or interventions in these studies would have differed in ways that would have impacted on the results, and therefore we should not assume a common effect size. Therefore, in these cases the RE model is more easily justified than the FE model. Additionally, the goal of this analysis is usually to generalize to a range of populations. Therefore, if one did make the argument that all the studies used an identical, narrowly defined population, then it would not be possible to extrapolate from this population to others, and the utility of the analysis would be limited.

If the number of studies is very small, then it may be impossible to estimate τ^2 with any precision. In this case, the FE model may be the only viable option. In effect, we would then be treating the included studies as the only studies of interest, and assuming that variation among them is absent or negligible for the parameter of interest.

3.2.2. Choice of model for sparse data

When the outcome of interest is relatively rare, few or no events may occur in one or both arms in several studies. Examples include an important but uncommon adverse event or

mortality in populations with low baseline risk. In these cases, the Normal approximation to the binomial distribution may not be appropriate, and commonly used meta-analysis methods may not yield correct confidence intervals (Sweeting, Sutton et al., 2004; Bradburn, Deeks et al., 2007). This situation occurs frequently when trials designed to test efficacy are pooled to estimate the rate of rare adverse events.

Such trials usually have smaller sample sizes than would be needed to establish effects on incidence of the rare events. We should recognize that the presence of a large number of such trials could result in a distorted estimate of harms because a substantial number of patients experiencing no events have been excluded from the analyses (Nissen and Wolski, 2007). Unless individual patient data are available for analysis, such trials are often excluded from pooled estimates of harms because a relative risk or odds ratio cannot be calculated.

Studies with no events in both arms may present particular problems, as some measures of treatment effect cannot be calculated in this case. If it is necessary to exclude them from the main analyses, they should be summarized qualitatively. Sensitivity analysis may be conducted to examine the effect on risk difference of including such trials (see Section 5.5 for detail).

For rare dichotomous outcomes, a random-effects model would provide a biased estimate of between-study variance. Ordinary logistic regression or the Mantel-Haenszel method (both FE methods) will provide a more robust estimate of combined effect, at the cost of disregarding the observed heterogeneity. These methods, or the Peto method (also FE), is preferable to the inverse-variance method and to some RE methods. Generally, the Peto method is preferred when the event rates are very low (< 1%) because it is least biased, and gives best CI, provided there is no substantial imbalance between groups. Bias is evident only in extreme imbalances (e.g. 8:1) and for large effects (e.g. OR <=0.2 or OR >=5) (Greenland and Salvan 1990, Bradburn, Deeks et al. 2007).

When a trial has no events in both arms, relative measures (e.g. OR and RR) are undefined. Some experts (e.g., Sweeting, Sutton et al. 2004, Bradburn, Deeks et al. 2007) consider studies with zeros to be non-informative and propose excluding them; others (e.g., Sankey, LA et al. 1996, Friedrich, Adhikari et al. 2007) consider including such studies to avoid biasing the results in the direction of a higher event rate. However, this can be resolved by presenting results on the risk scale rather than on the log-odds scale, taking account of trials with zero events just in this presentation step (Lane 2013). So when relative measures are used, exclude studies without any events from the main analyses. If the sample sizes are small, including or excluding such studies does not change the summary effect size (OR or RR) appreciably because they receive very small weight in the analysis. When analysing rare-event data, always perform sensitivity analyses to look for changes in the magnitude or the variance of the summary effect.

References

Altman DG, Deeks JJ (2002). Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology* **2**:3.

Bradburn MJ, Deeks JJ, Berlin JA, Localio AR (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* **26:**53–77.

Cates CJ (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Medical Research Methodology* **2**:1.

Collins R, Peto R, MacMahon S, Godwin J, Qizilbash N, Hebert P, Eberlein KA, Taylor JO, Hennekens CH, Fiebach NH (1990). Blood pressure, stroke, and coronary heart disease: Part 2, short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *The Lancet* **335** (8693):827–838.

Friedrich J, Adhikari N, Beyenne J (2007). Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Medical Research Methodology* **7**: 5.

Greenland S, Salvan A (1990). Bias in the one-step method for pooling study results. *Statistics in Medicine* **9**:247–252.

Hutton JL (2000). Number needed to treat: properties and problems. *Journal of the Royal Statistical Society A* **163**:381–402.

Lane PW (2013). Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research* **22**:117–132.

Nissen SE, Wolski K (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine* **356**: 2457–2471.

Parmar MK, Torri V, Stewart L (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* **17**:2815–2834.

Quartey GK, Blackman N, Williams M (2007). *Use of Number Needed to Treat in Clinical Trials*. Technical Review Document, Drug Discovery Sciences, GSK.

Sankey W, L, et al. (1996). An assessment of the use of the continuity correction for sparse data in meta-analysis. *Communications in Statistics – Simulation and Computation* **25**:1031–1056.

Sweeting MJ, Sutton AJ, Lambert PC (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* **23**:1351–1375.

4. GRAPHICS AND SOFTWARE

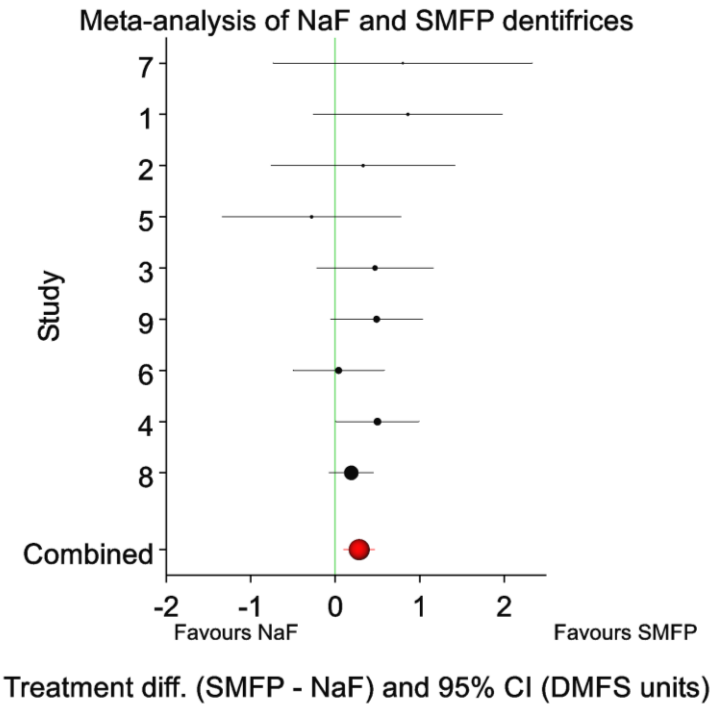
4.1. Graphical methods

Graphical methods are the most effective way of summarizing the information in a meta-analysis and communicating the results. Tables and text are needed to provide the background and detail to an analysis, and to discuss the interpretation, but a good set of graphical displays will provide the key messages. Because graphs will be more focused on than text by the readers of a report, it is essential that that they should be constructed to avoid distorting or obscuring the information.

4.1.1. Interval plots

There were two examples of interval plots in Section 2, one illustrating sensitivity analysis and the other summarizing the data and combined results of a meta-analysis. This second use is often called a forest plot, and is the most commonly used graphical display in meta-analysis. Figure 4.1 shows it again.

Figure 4.1. Forest plot of studies of dentifrice efficacy.



The studies are ordered here so that the largest studies are at the bottom, near the combined estimate: it helps understanding if a sensible ordering like this is imposed, rather than presenting them in an arbitrary order (e.g. alphabetically by Study name) or

ordered by a relatively unimportant variable (e.g. date of Study).

It has become conventional to present 95% confidence intervals around the estimates in a forest plot, rather than any other indications of variability such as standard errors. However, for clarity, it should always be stated what the intervals represent.

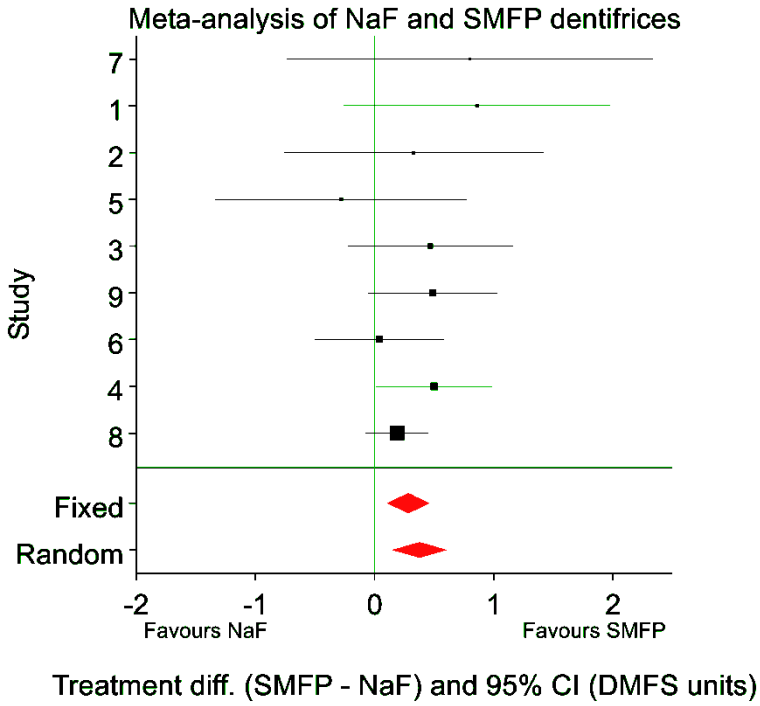
The estimates are usually plotted with a symbol whose area is proportional to the weight of the estimate in the meta-analysis. This is intended to counteract the visual effect of the longer intervals appearing more prominent than the shorter ones, when in fact they make less contribution to the combined estimate. Most forest plots seem now to have square symbols, though circles are sometimes seen, as here, because the "bubble plot" from which it derives usually has circles; the more conventional form is shown next. Note that it is the area of the symbol, not the width or diameter, which is used to indicate weight, because the visual impression of importance is roughly proportional to the area. The combined estimate is usually presented at the bottom of the plot as a diamond, whose horizontal points indicate the 95% confidence limits. Ideally, its area should be in the same proportion to weight as for the individual estimates, as in Figure 4.1; i.e., its area should be the sum of the areas of the component studies. However, this is rarely done in practice because the diamonds drawn by standard software tend to have fixed height, and symbols for individual studies would be too small to be seen clearly. Alternatively, the same symbol can be used as for the individual estimates, relying on labelling and spacing to make clear the distinction, as in Figure 4.1.

If several meta-analyses are carried out on the same data, for example using both fixed- and random-effect models, several alternative combined estimates can be presented at the bottom of the graph. It is also common to see a series of subgroups analysed and displayed in a single graph, with combined estimates shown for each subgroup as well as overall.

Most specialized meta-analysis packages can draw forest plots (see Section 4.2), but general statistical packages don't usually provide for them.

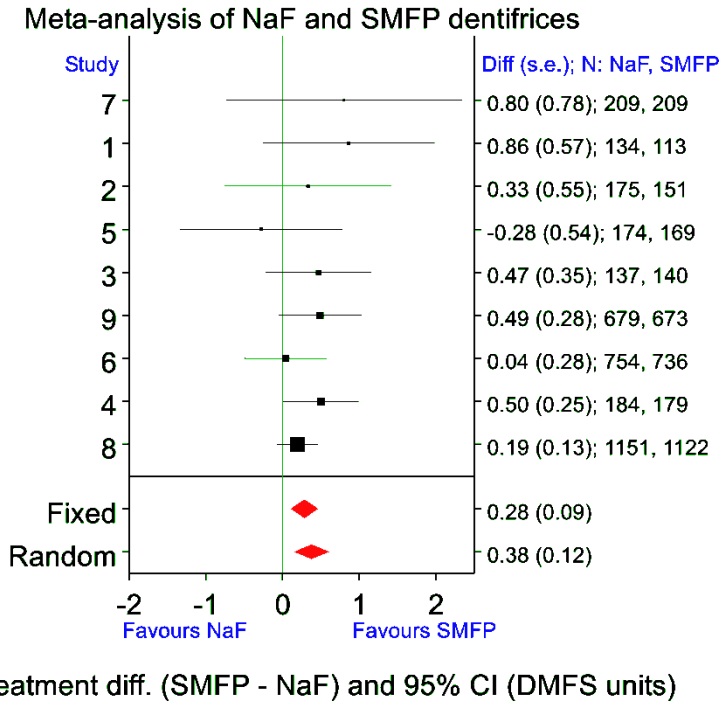
Figure 4.2 shows a second version of this forest plot, with square and diamond symbols, and showing fixed and random summaries. The summaries have been marked off as separate from the study information by placing them in a graphical margin.

Figure 4.2. Forest plot with contrasting symbols and two marginal summaries.



A further common addition to the forest plot is the numerical information about the component trials and the combined estimates. This can be useful to provide extra context – particularly the numbers of patients involved in each of the trials. Figure 4.3 shows an example, with the estimated differences and SEDs on the right-hand axis, together with numbers of patients. The simplest way to add this information using general graphical software is as labels for an additional axis on the graph; but specialist software for meta-analysis should provide options to add the information. There is a discussion of features in the forest plot and recommendations in Anzures-Cabrera & Higgins (2010).

Figure 4.3. Forest plot with additional annotation.

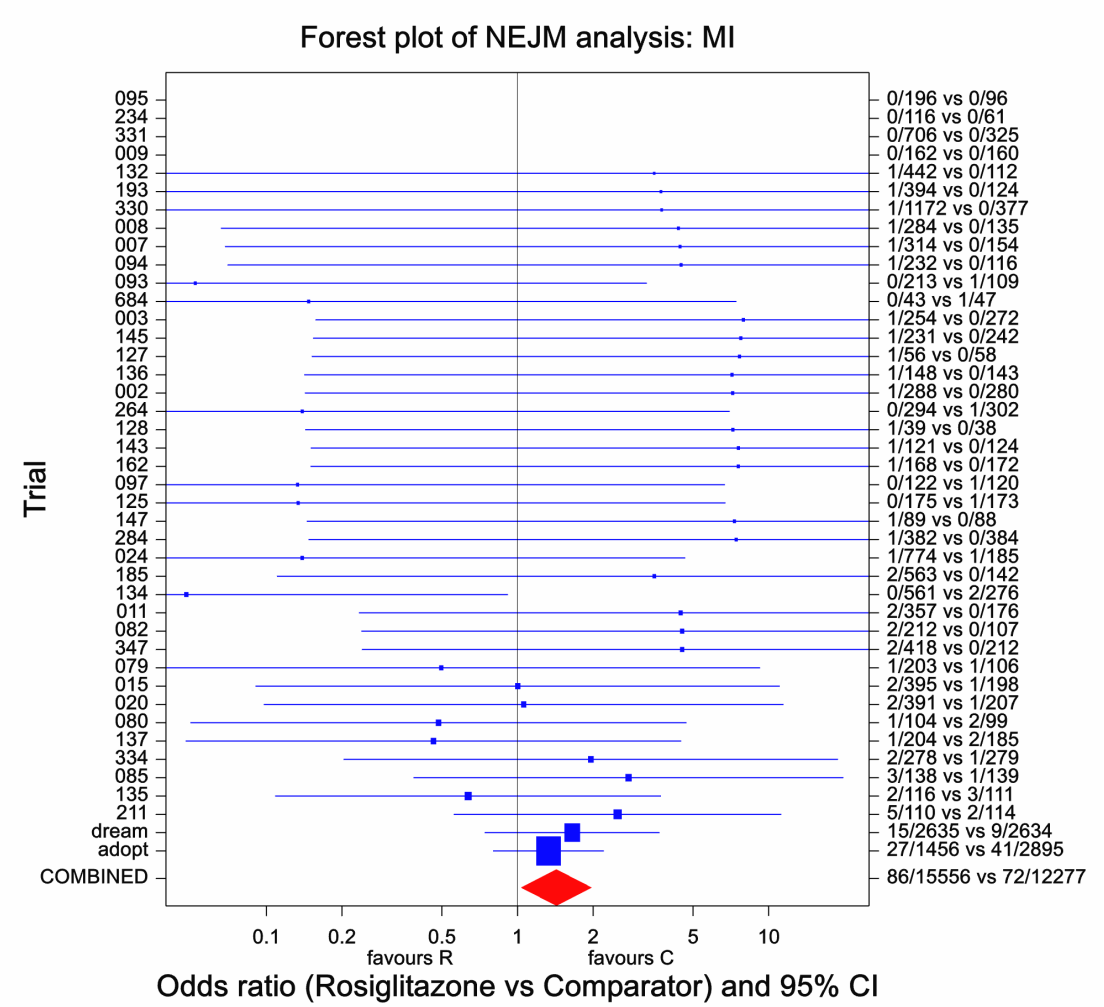


The same form of display can be used effectively to display meta-analysis carried out on any scale and with any method, whether based on summary statistics or on patient-level data. For example, Figure 4.4 shows a display of the analysis of Rosiglitazone trials published in the NEJM, investigating a possible connection with myocardial infarction (MI). The summary information from 42 trials was analysed using the Peto method (see Section 5.4) which is suitable for combining information on incidence of rare events on the odds-ratio scale.

The figure shows many features of the data that would otherwise have to be gleaned by reading much text or studying several tables. The blank space at the top demonstrates that there were four trials with no events at all, and which make no contribution to the combined estimate of odds ratio (though there were actually 10 such trials, as discussed in Section 5.5). The very wide confidence intervals for most trials indicate how little precision is associated with the estimates based on each of these. Some of the intervals, indeed, extend even beyond the wide range of about 0.05 to 20 on the odds-ratio scale, and are clearly truncated to avoid cramping the information from the larger trials within too wide a range. Notice that the x-axis has a log scale: this is usually appropriate when an analysis uses a log transformation, as here, because the confidence intervals will then be symmetric. It is also usually best to label the axis with values on the untransformed scale, as this is more readily understood (though still representing an odds ratio, which is a difficult concept for many non-statisticians). The size of the bottom two squares, together with their associated short intervals and the large numbers of patients on the

right-hand axis, make it clear that these two large trials dominate the analysis. Finally, the diamond representing the combined estimate shows that it has just achieved statistical significance at the 5% level.

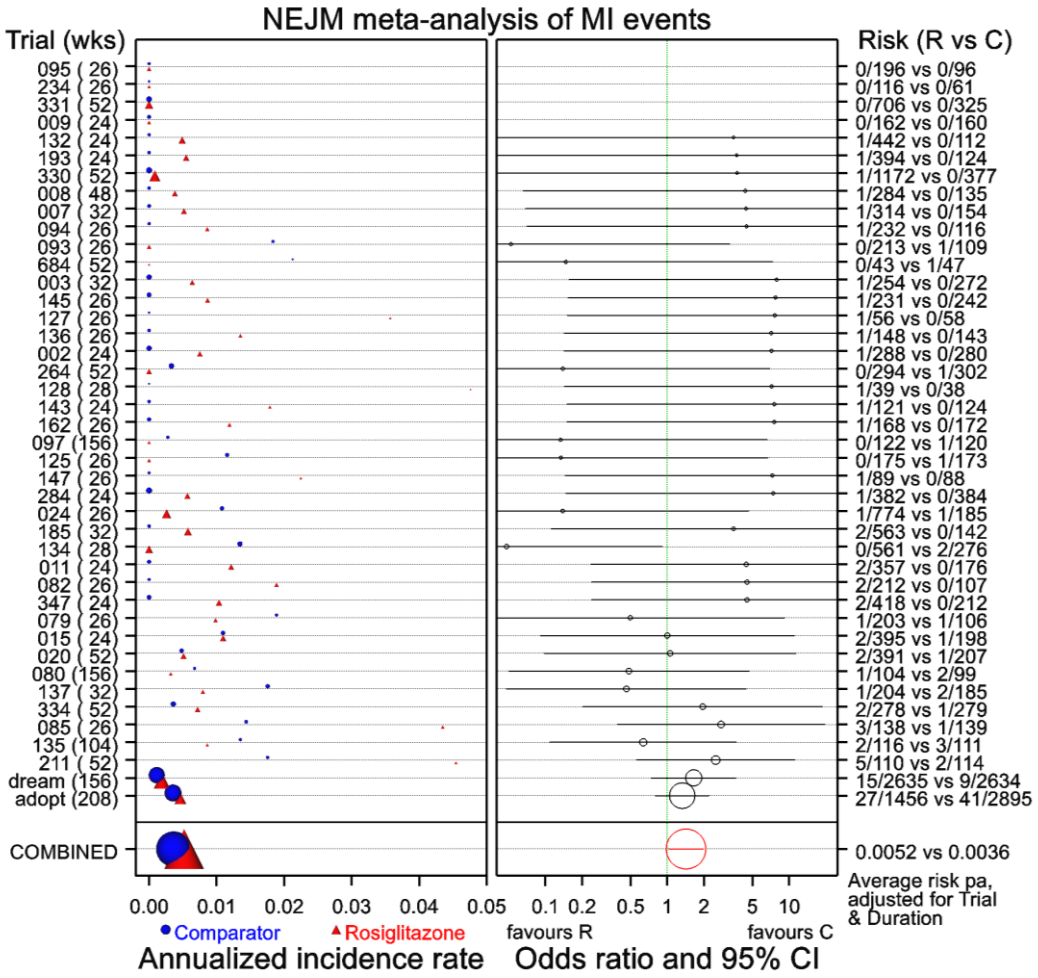
Figure 4.4. Forest plot of the Rosiglitazone analysis.



One major problem with this display, and with a similar presentation of any meta-analysis carried out on a transformed scale, is that there is no interpretation of the results on the natural scale of risk. Indeed, the label on the right-hand axis against the combine estimate shows just the pooled estimates of risk in the two treatment groups, unadjusted for imbalance: a naïve interpretation of these contrasts with the estimate from the meta-analysis, as the risk is apparently higher for Comparator (72/12277=0.0059) than for Rosiglitazone (86/15556=0.0055) whereas the combined estimate (red diamond) favours the comparator. We shall see in Section 5.5 that this is an example of Simpson's Paradox. Less confusing would be a summary of the combined estimate based on the model, the details of which we will also look at later.

A second problem with the interpretation of the combined estimate is that the trials were of very different length. This is taken account of in the analysis on the log-odds scale, but needs explicit scaling on the risk scale. Figure 4.5 shows an improved display, with an additional panel displaying the raw risks in each study graphically, and a pair of predicted risks averaged over the mix of populations in the Rosiglitazone arms, expressing the annualized incidence rate.

Figure 4.5. Joint dot plot and interval plot of the Rosiglitazone analysis.

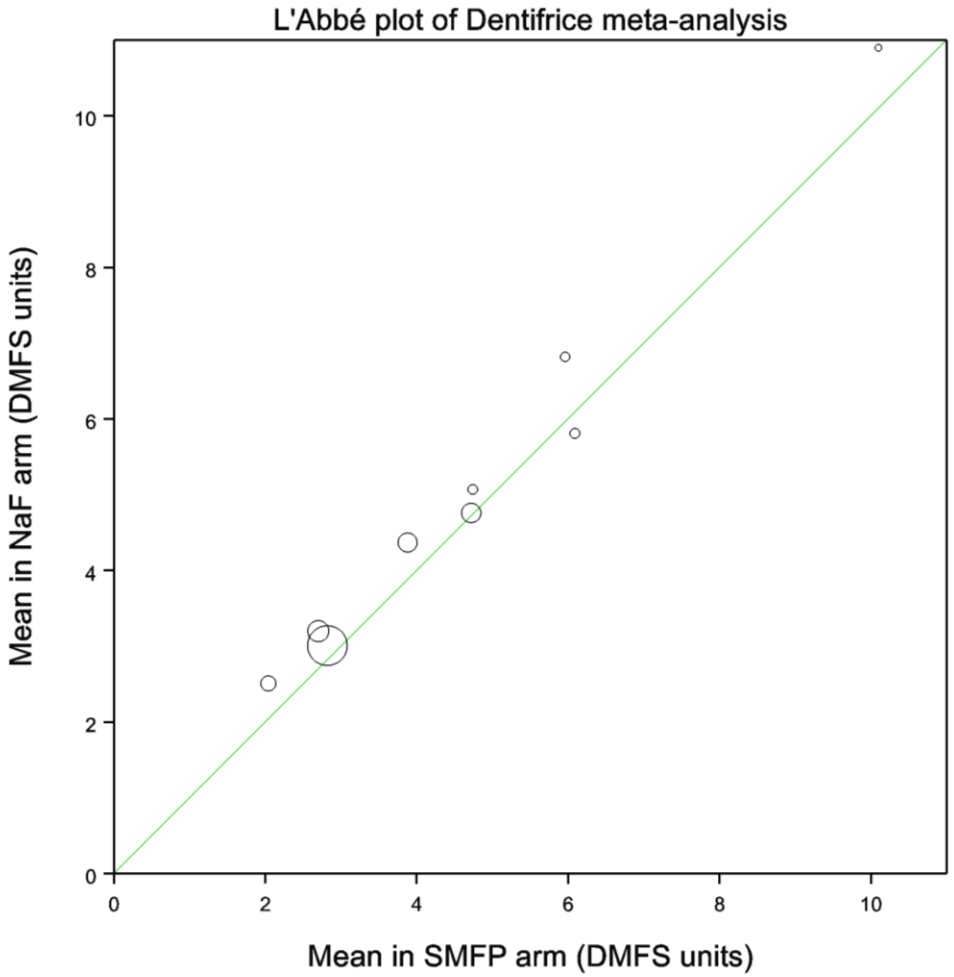


4.1.2. Scatter plots

One of the simplest plots to illustrate a meta-analysis is a scatter plot of the contributing summaries from each trial, one arm against the other. It helps to add a reference line at equality, to allow assessment of the scatter either side, and to vary the symbol size according to the number of subjects, or to the weight in a meta-analysis (i.e. a bubble

plot). Figure 4.6 shows an example for the dentifrice analysis.

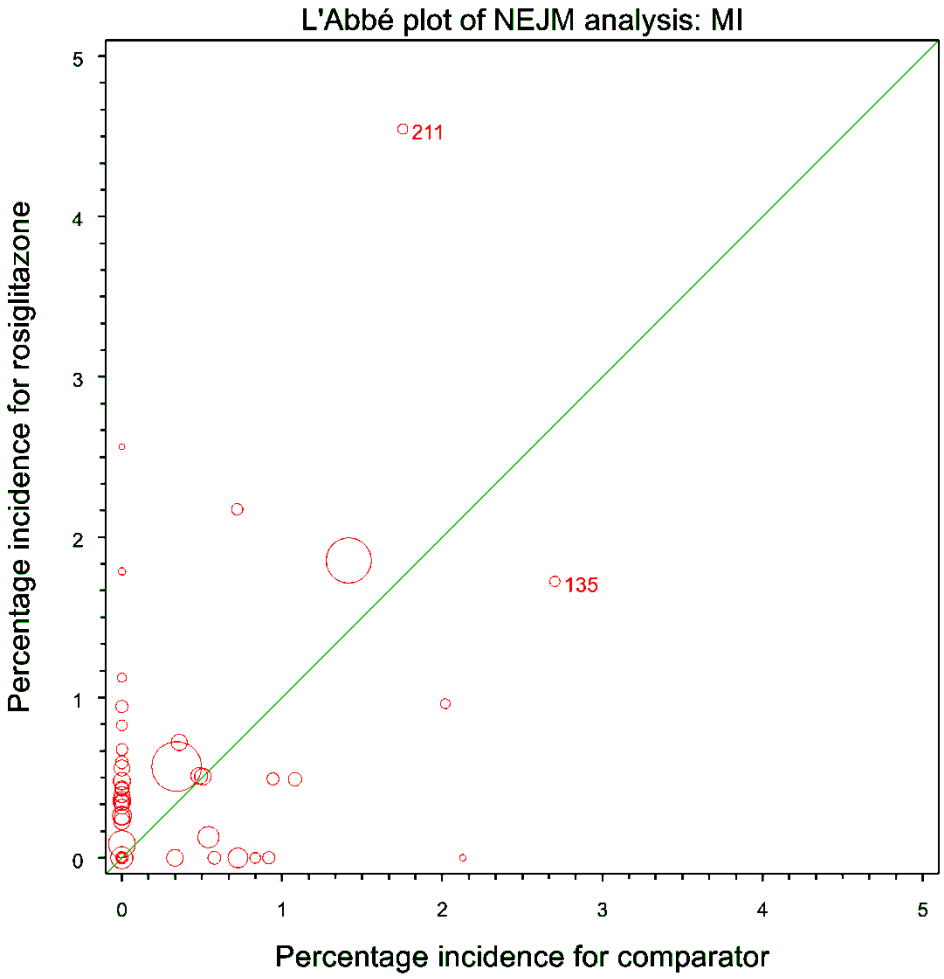
Figure 4.6. L'Abbé plot of the dentifrice analysis.



This demonstrates that nearly all the studies are on one side of the line, favouring SMFP; it also highlights that one study is on patients with much higher average scores of decay.

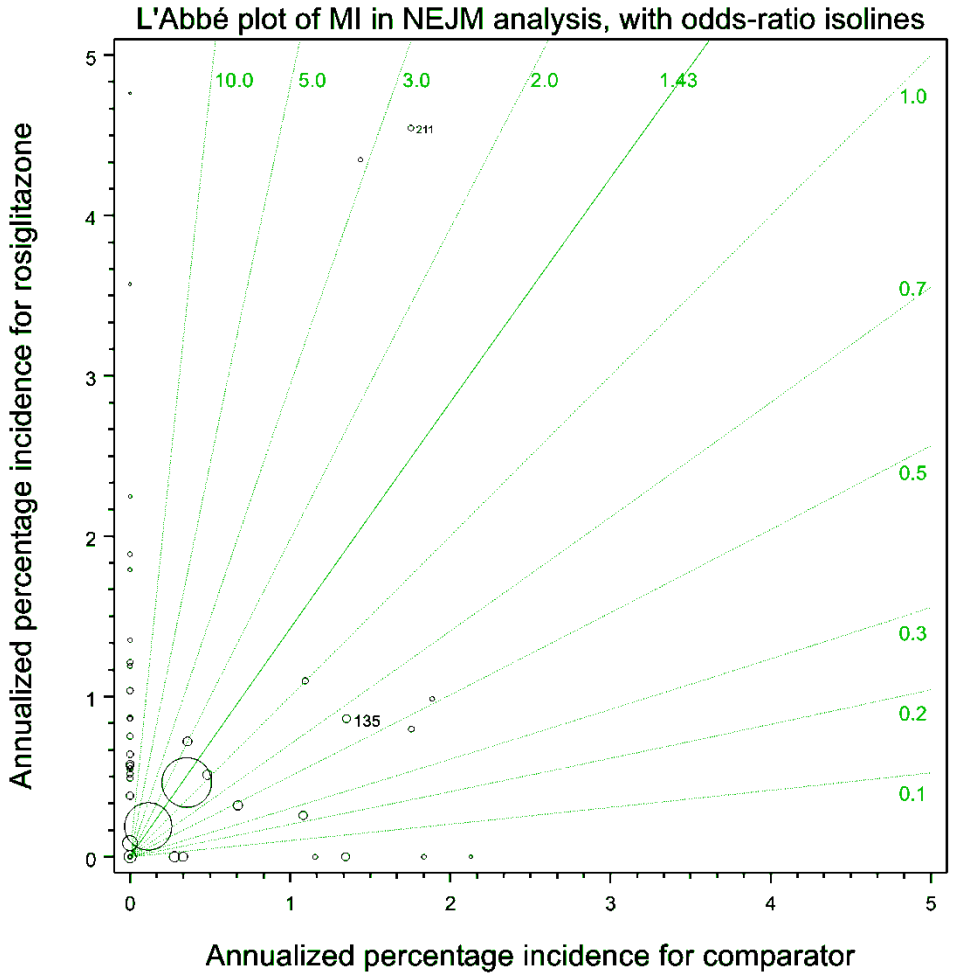
This type of plot was first used by L'Abbé et al. (1987), and so is called a L'Abbé plot. It is more often used to summarize analysis of risks; Figure 4.7 shows an example for the Rosiglitazone meta-analysis.

Figure 4.7. L'Abbé plot of the Rosiglitazone analysis.



This display emphasizes the large number of trials with no events in one or both arms, and the wide scatter either side of the reference line. The two studies with the highest observed incidence (because of the population of patients studied) are individually labelled. However, the information here is distorted by the fact that the duration of the trials varies considerably: from 24 to 208 weeks. So if the incidence of MI events is expected to be constant for a trial population during the course of the trial (rather than having an initial rise due to treatment and then falling, for example) the longer trials are bound to have higher incidence rates. This can be adjusted for by annualizing the rates. This has been done in Figure 4.8, which also includes "iso-lines" of the odds-ratio – producing a type of contour plot.

Figure 4.8. L'Abbé plot of an annualized Rosiglitazone analysis with odds-ratio iso-lines.



A second common use of the scatterplot is in the assessment of publication bias. A plot of 'study size' associated with each contributing estimate, against the estimate itself, is referred to as a 'funnel plot'. The 'study size' can be expressed as sample size, SE, or 1/variance; Sterne & Egger (2001) recommend the use of SE. The plot can indicate a lack of studies with low precision and non-significant results, a situation which arises with many collections of published (rather than in-house) trials because of the tendency not to publish the results of trials showing no significance. This important aspect of meta-analysis is covered in more detail in Section 7.2. Figure 4.9 shows an example for the Rosiglitazone data, showing a reference line at OR=1.43, which was the reported combined estimate. The reference "funnel" shows pointwise 95% confidence limits about the reference line, allowing a visual check of the distribution of the treatment effects from each study around the combined estimate. Study 134 is just outside the funnel, and is labelled individually; but to have one study from 38 outside the 95% range is quite

reasonable (the four studies with no events are not represented). It is conventional to reverse the y-axis in this diagram so that the point of the funnel (corresponding to studies with infinite sample size and zero SE) is at the top. Figure 4.10 shows an alternative version using precision rather than SE on the y-axis; the display is more cramped, and the reference lines curved rather than straight.

Figure 4.9. Funnel plot of the Rosiglitazone analysis.

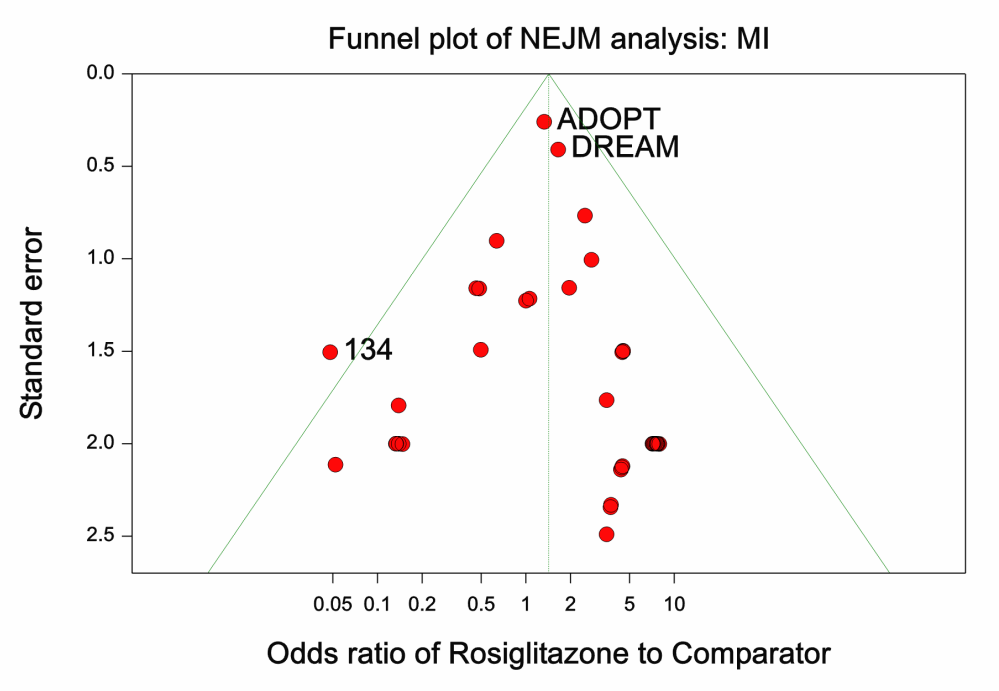
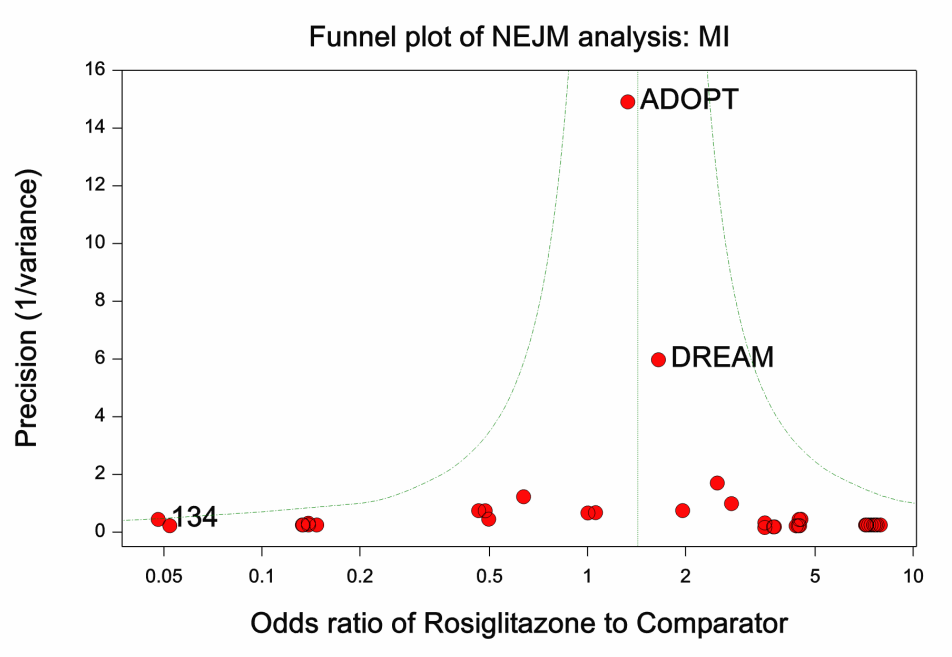


Figure 4.10. Funnel plot of precision vs estimate.



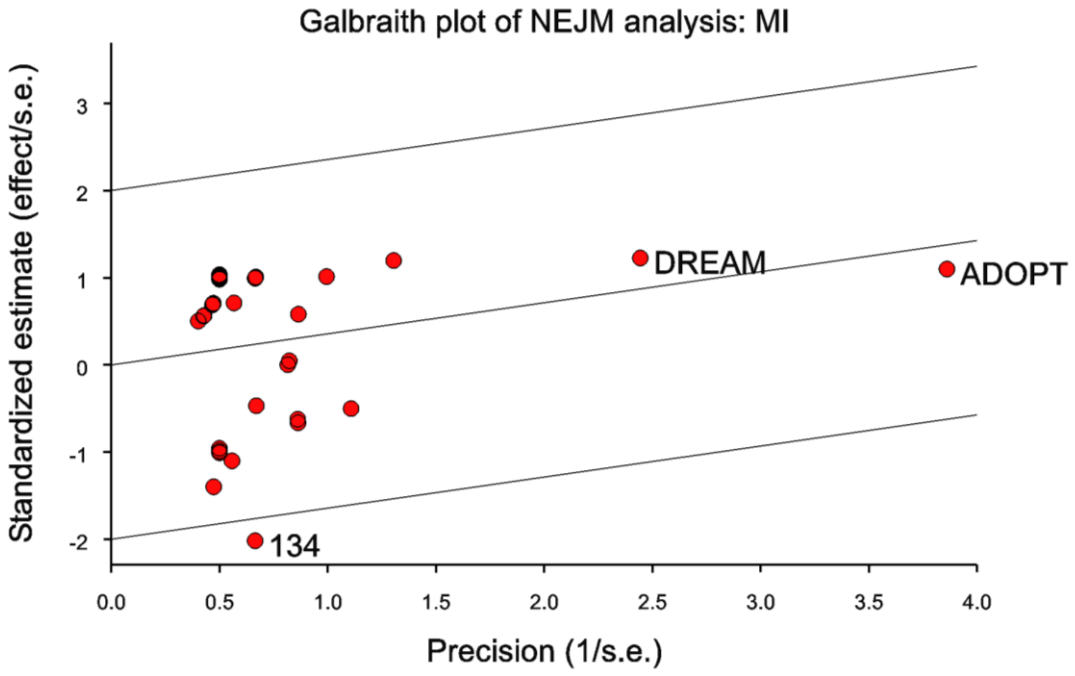
The two large studies form the apex of the 'funnel' and the others are all scattered either side of the combined estimate. There is no indication here of publication bias, which is not surprising because the GSK Clinical Trial Register was searched to find relevant trials, and this register aims to record all trials regardless of outcome. There is indication of some skewness of the distribution on this log-odds scale, which is the appropriate scale on which to assess this. However, there are many potential causes underlying asymmetry in this type of plot: see Section 7.5.1 of Sutton et al (2000) for more information.

Both these types of scatterplots are straightforward to draw with standard software, though the sizing of the bubbles in the l'Abbé plot may require careful programming.

4.1.3. Radial plots

There is another standard way to present information from meta-analysis on an odds-ratio scale. Starting from a funnel plot using 1/SE as the measure of 'study size', the estimates are standardized (divided by their SEs) and the axes are interchanged. This gives what is called a radial plot, or Galbraith plot (Galbraith, 1988). Figure 4.11 shows the main part of the plot for the Rosiglitazone meta-analysis.

Figure 4.11. Galbraith plot of the Rosiglitazone analysis.

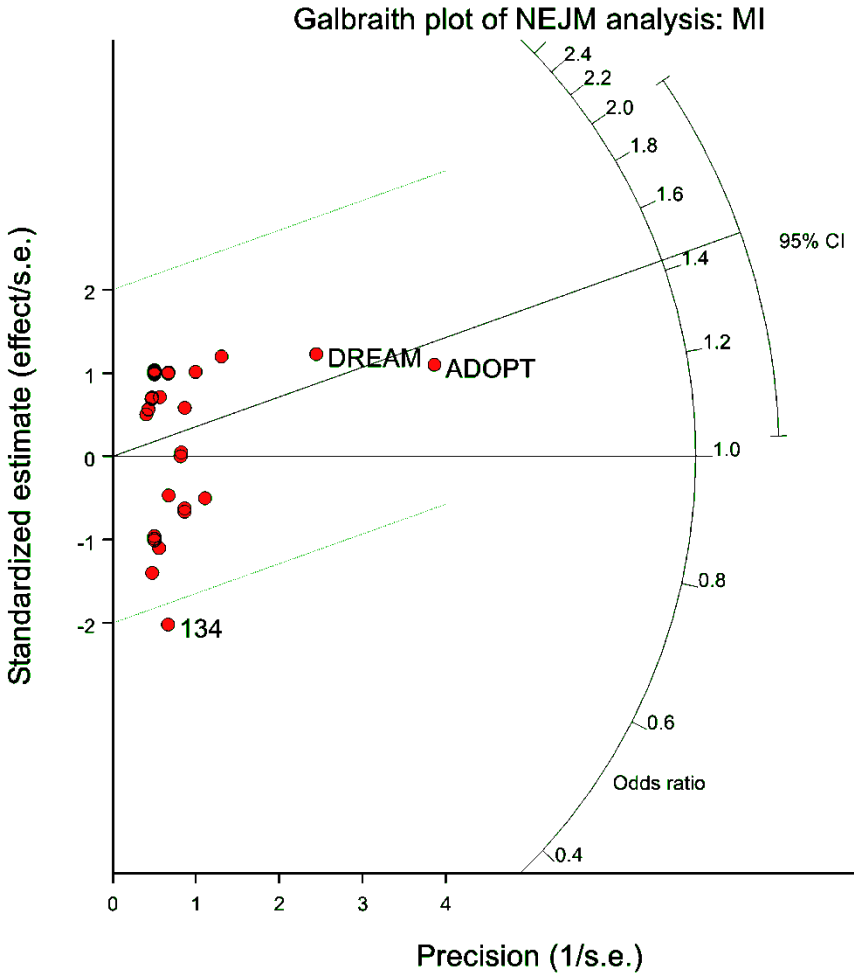


The plot takes advantage of the property that the combined estimate is the slope of an

unweighted regression line fitted through the origin of the scatterplot (the points have already been effectively weighted by scaling by the SE). It clearly demonstrates here how strongly the combined estimate depends on the two large trials. The plot also includes two reference lines that give the 95% confidence limits for the standardized estimates under the model; this allows visual assessment of heterogeneity.

The full form of the radial plot includes a circular (or sometimes vertical) axis to calibrate the slope (i.e. the combined estimate) with its confidence interval (Figure 4.12). This also adds to the visual appreciation of how the individual study estimates contribute. This axis is again best labelled on the odds-ratio scale rather than the log-odds scale on which it is based. The addition of the circular axis requires detailed programming unless a specialized meta-analysis package is available (see Section 4.2); the basic plot as above is straightforward to draw with standard software.

Figure 4.12. Full radial plot of the Rosiglitazone analysis.



4.2. Software

Most meta-analysis techniques are fairly straightforward and can be programmed in any general language by a competent programmer: for example, the Excel spreadsheets that are available for use in the practical sessions. Several general statistical systems provide a range of techniques already programmed in the form of macros or procedures, but not usually as items in a menu interface. This section describes what is available in SAS, S-Plus, Stata, R and GenStat. The simplest way to do the calculations is to use a special-purpose package designed specifically for meta-analysis. The main contenders are Comprehensive Meta-Analysis (CMA), MetAnalysis (MA), MetaWin (MW), MIX (developed as a PhD project at Kitasato University), Review Manager (RM) and Open Meta (OM).

Patient-level analysis can be handled with standard statistical software, because the facilities required are the same as for analysing multi-centre trials.

4.2.1. Special-purpose packages

A comparative review is given in Bax et al. (2007), though this is now rather out of date (and does not include OM which was developed since 2007). They found and mention many other packages, but exclude them on the basis of not having been recently updated (in the previous five years) or having no menu interface. Here is a brief summary of each of those reviewed by Bax, taking information from the review and from our own experience. The tables below summarize the methods provided, and graphical capability. We have omitted WEasyMA (included by Bax et al), because it is apparently no longer developed. Prices are as shown on the Internet in June 2015.

Comprehensive Meta-Analysis ($895 per year for Professional version) http://www.meta-analysis.com/
This is the most expensive, and has the highest profile on the Internet. It has comprehensive numerical options and output, providing reasonable graphics. Data can be entered into its spreadsheet by typing or pasting, in many forms (e.g. proportions, *r* out of *n*, risk ratios), but it cannot import data files such as text, Excel or SAS. The tutorial and manual are extensive, but it has no in-program help. With a binary response, it adds 0.5 to all cells if any are zero: a feature which cannot be switched off.

MetAnalysis (£84) Leandro (2005)
This is not sold separately, and has no website, but comes as a bonus feature of a book by Leandro (2005). It only handles summary analysis for binary response. Data cannot be pasted or imported, so must be entered manually. There is no in-program help, nor even a website, but the book serves as a manual.

MetaWin 2.1 ($60 academic users only) http://www.metawinsoft.com/
This is accompanied by a comprehensive manual in the form of a book. It has a

spreadsheet interface and can import various types of datafile. There is extensive in-program help. One downside is that data-ranges have to be selected each time an analysis is changed. Unusually, calculations are based on t- rather than Normal distributions, and bootstrap methods are available for confidence intervals.

MIX 2.0 ($210) http://www.meta-analysis-made-easy.com/
This is the most recently developed package, and is designed as an Excel add-on. It has detailed numerical options, and educational features like built-in data sets corresponding to those in a number of books, and extensive tutor functions. Data can be manually entered or imported from text or Excel files. It provides either constant or treatment-arm "continuity correction" for a rare binary response.

RevMan 5.1 (free for private and academic use, €850 for commercial use)
http://ims.cochrane.org/revman/download
This was developed by and for the Cochrane Collaboration. It has extensive features for collaborative management of systematic reviews, and analysis can't actually be done at all without creating a review structure. Data importation and pasting is also limited, so it takes longer to get started than with other software. The help resources are extremely thorough.

Analysis capability (from Bax et al)
All the packages provide the DerSimonian Laird method for random-effects, and inverse-variance, Mantel-Haenszel and Peto for fixed-effect models. MA does not handle continuous data; the others analyse mean difference (except MW) and also Hedges G; CMA and MIX analyses Cohen's D. For binary data, they all work with risk differences, risk ratios and odds ratios (except for MA, which doesn't handle risk ratios). All except RM offer one or more methods for analysing small-study effects or publication bias.

Table 4.1. Graphics capability of meta-analysis packages (from Bax et al).

	Forest	Funnel	Galbraith	L'Abbé	Sensitivity
CMA	Y; prop, annot	Y; se^{-1},se	N	N	Y
MA	Y; annot	Y; N	Y; radial	Y	N
MW	Y	Y; Var, N	Y	N	N
MIX	Y; prop, annot	Y; se^{-1},se, N, P	Y	Y	Y
RM	Y; prop, annot	Y; se^{-1}	N	N	N

Forest plot: only some packages allow row annotation (annot) or proportional symbol sizing (prop)
Funnel plot: several alternative measures of precision are offered
Galbraith plot: only MA adds the radial information

Open Meta (free, open-source software) http://www.cebm.brown.edu/open_meta/
This is developed by the Center for Evidence-based Medicine at the Brown School of Public Health, and is supported by AHRQ (Agency for Healthcare Research and Quality).

It has an extensive choice of FE and RE methods, including meta-regression and subgrouping. It is not clear from the website whether datasets can be imported. Graphics includes at least Forest plots.

4.2.2. General statistics packages

GenStat

The META procedure was written for Stephen Senn's courses on meta-analysis. It covers the methods in Chapter 4 of Whitehead (2002) based on summary estimates, for both fixed-effect and random-effects models. Either maximum-likelihood or REML can be used for estimation of random effects. It includes tests for heterogeneity and can draw Forest plots (with variable symbol-size if requested, but not row annotation) and Galbraith plots. GenStat provides spreadsheet data-entry, and can import most types of datafile.

R

The 'rmeta' package provides a range of functions for meta-analysis, providing the Mantel-Haenszel method for fixed effects, and the DerSimonian-Laird method for random effects. There are functions to draw Forest and Funnel plots.

http://cran.r-project.org/web/packages/rmeta/index.html

There is also a 'meta' package:

http://cran.r-project.org/web/packages/meta/index.html

and the metafor package will fit fixed and random effects models via the general linear (mixed) model and produce most standard plots:

http://cran.r-project.org/web/packages/metafor/index.html

There are also some relevant functions in the epiR package:

http://cran.r-project.org/web/packages/epiR/index.html

The MiMa function written by Wolfgang Viechtbauer apparently works with R as well as S-Plus. R can import most types of datafile.

SAS

SAS has no intrinsic features for meta-analysis. Stephen Senn has produced a suite of nine detailed macros, which can be downloaded from:

http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html

David Wilson has also produced some simple macros, which can be downloaded from

http://mason.gmu.edu/~dwilsonb/ma.html

We have shown examples of SAS commands for many of the methods, which can be used if you are prepared to program an analysis. SAS's main strength is its flexibility for data-handling, and you can perform a number of meta-analyses relatively easily using standard procedures. For example, the FREQ Procedure provides Mantel-Haenszel analysis, and also some exact methods. Plotting is difficult, but there is an example of a Forest plot in the BDS Graphics Catalogue. SAS can import only text and Excel datafiles.

S-Plus

S-Plus also has no intrinsic features, but functions have been written. These include MiMa, written by Wolfgang Viechtbauer, from

http://www.wvbauer.com/downloads.html

Graphics is generally easier to program in S-Plus than in SAS and there is an example of drawing a Forest plot in the BDS Graphics Catalogue. S-Plus can import most types of datafile.

Stata

Stata provides a very wide range of user-written commands for meta-analysis, which can be installed over the Internet from within the system. The 'metan' command covers inverse-variance, Mantel-Haenszel, Peto and DerSimonian-Laird. It also draws Forest plots (with or without annotation and variable symbol-size). Additional commands provide Funnel plots ('funnel') and L'Abbé plots ('labbe'). The 'metan' command adds 0.5 to cells automatically when there are zero cells. There are also commands for meta-regression ('metareg'), cumulative meta-analysis ('metacum'), studying influence ('metainf'), bias-testing ('metabias'). Stata can import most types of datafile.

WinBUGS/OpenBUGS/JAGS

Often used for Bayes meta-analysis, but requires some coding or editing of other's code; easiest used from one of the interfaces to standared software packages, eg R2WinBUGS.

http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml
http://cran.r-project.org/web/packages/R2WinBUGS/index.html
http://www.openbugs.info/w/
http://www-fis.iarc.fr/~martyn/software/jags/

4.2.3. Nonlinear mixed-effects packages

NONMEM 7

NONMEM is a nonlinear mixed-effects modelling tool used in population analysis. It tool includes many different estimation methods such as Monte-Carlo expectation-maximization and Markov-Chain Monte-Carlo Bayesian methods in addition to the classical likelihood approach.

References

Anzures-Cabrera J, Higgins JPT (2010) Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods* (2010) **1**:66–80.

Bax L, Yu L-M, Ikeda N, Moons KGM (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology* **7**:40.

Galbraith RF (1988). A note on graphical presentation of estimated odds ratios from several trials. *Statistics in Medicine* **7**:889–894.

L'Abbé KA, Detsky AS, O'Rourke K (1987). Meta-analysis in clinical research. *Annals of Internal Medicine* **107**:224–233.

Leandro G (2005). *Meta-analysis in Medical Research*. Blackwell, BMJ Books.

Sterne JAC, Egger M (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* **54**:1046–1045.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000). *Methods for Meta-analysis in Medical Research*. Wiley: Chichester.

Whitehead A (2002). *Meta-analysis of Controlled Clinical Trials*. Chichester: Wiley.

5. FIXED-EFFECTS APPROACHES

5.1. Continuous response

Section 2 illustrated the meta-analysis of treatment differences estimated from studies with a continuous endpoint. It used the inverse-variance approach to weight the differences, which is the most common way used to combine the information. The calculation is straightforward, and can easily be achieved using standard statistical software. The combined estimate is just the estimate from a weighted ANOVA in which the variance is fixed at 1.0. For example, in SAS, say the estimates and SEs from each study have been calculated and stored in a dataset *D* which has one row per study:

Study	Estimate	SE
1	5.21	0.75
2	6.04	0.44
3 …		

All that is now needed is to add to D a weight variable, *Weight* say, calculated as the reciprocal of the squared s.e.s, and then use:

```
proc mixed data=d noprofile;
  model estimate = /solution;
  weight weight;
  parms (1) /noiter;
  run;
```

The PARMS statement is needed here to fix the variance at 1.0 so that the SE of the combined estimate is formed correctly, and the NOPROFILE parameter is also needed to stop SAS from profiling the variance parameter. Unfortunately, the variance cannot be fixed in the GLM procedure, so the MIXED procedure has to be used, even though there are no random effects apart from the usual residual term.

One potential difficulty in this situation is when the collected studies do not all use the same endpoint; then, it may be appropriate to work with standardized treatment differences.

When patient-level data are available, all the data need to be assembled to allow a model to be fitted as outlined in Section 2.5. This can be done using the same approach as would be used for estimating the treatment difference in a single study, with the inclusion of an extra categorical term in the model to account for the trial-to-trial differences. So, we might construct a dataset *C*, say, for SAS, which has one row for each subject in all the chosen trials, and stores the endpoint together with identifiers of the associated treatment and study.

Study	Subject	Treatment	Response
1	1001	active	4.53
1	1002	control	4.21
1 …			

The meta-analysis can then be carried out with statements like

```
proc glm data=c;
  class Treatment Study;
  model Response = Treatment Study …
  repeated /group=study;
  run;
```

where the MODEL statement includes the names of all the covariates that are available and considered suitable for inclusion.

Patient-level analyses are generally much harder and more time-consuming to carry out than summary-level analyses. One reason is the major step of bringing together the patient-level data from all the identified studies. It is likely that definitions of variables and recording methods will vary from study to study, even if all studies were run by the same organization. More seriously still, the choice of which covariates to include in the analysis will be constrained by the availability of the information across the studies. Most clinical trial records should include common candidates such as Sex, Age and Baseline; but other variables that may be relevant to the particular application, such as smoking status or time since first diagnosis, may be recorded only in some trials. A decision then has to be made whether to compromise on the variables included, in order to have information that is as complete as possible, or to compromise on the studies included, in order to have a model that is as flexible as possible. There are general statistical methods of imputing information for missing variables, based on models involving the other variables that are available; but these should be used cautiously, and cannot be relied on when large proportions of information are missing.

Two effects should always be present in a patient-level analysis: one to take account of the differences between studies, and the other to represent the effect of the treatments being investigated. The interaction between these two covariates is of great interest in a meta-analysis, and represents heterogeneity between the studies.

When combining information for randomized trials, there is a general level of protection from the effects of covariates afforded by the randomization process. We expect that on average the distribution of the values of any covariate will be approximately the same in each treatment group, so a linear analysis excluding that covariate should not be greatly biased. Exclusion will, however, tend to increase the error variance in an analysis assuming a Normal distribution for response. For a non-Normal analysis such as survival analysis or logistic regression, error variance is not an issue, but the effect of excluding an important covariate can be to bias the treatment effect. In such analyses, if the events are rare, individual covariate values can become important, but the analysis is correspondingly hard to interpret because effects of covariates and treatments may be confounded.

A second cause of difficulty in patient-level analysis is the analysis itself. Statistical

software can handle large datasets, but it can nonetheless be awkward and time-consuming to fit models to the many thousands of patient records that may be involved. There can be particular difficulty when looking at rare events, because some methods that use asymptotic results may not behave well. For example, in a logistic regression of a serious adverse event, there may be few instances of the event even across the whole set of trials, while it is generally necessary for the model to be at least complex enough to account for study and treatment differences.

5.2. Binary response: risk difference

The method for combining information on a binary endpoint is essentially the same as that for a continuous endpoint, once the individual estimates of treatment difference and their SEs have been calculated. However, there are several alternative measures of difference, as we saw in Section 3. We will illustrate these, and the calculations involved, using a well-known meta-analysis of seven RCTs (Table 5.1) of the effect of aspirin when given to heart-attack patients (Fleiss & Gross, 1991).

Table 5.1. Outcome data from seven aspirin trials.

	Aspirin		Placebo	
Trial	**patients**	**deaths**	**patients**	**deaths**
MRC-1	615	49	624	67
CDP	758	44	771	64
MRC-2	832	102	850	126
GASP	317	32	309	38
PARIS	810	85	406	52
AMIS	2267	246	2257	219
ISIS-2	8587	1570	8600	1720

A simple pooling of the information, ignoring trial differences, gives the impression that treatment with aspirin slightly reduces the risk of death from 16.5% for the Placebo patients (2,286 out of 13,817) to 15.0% for the Aspirin patients (2,128 out of 14,186). This simple pooling approach runs the risk of confounding the treatment effect with differences between the trials – leading to the effect known as 'Simpson's Paradox'. If the underlying risks are different in the different trials, for example because they represent different patient populations, then imbalance between the treatments leads to those differences in underlying risks being confounded with the treatment effect. With the trials above, there is little chance of this because there are similar numbers of patients receiving the two treatments in each trial, except for one, the PARIS trial, where twice as many received Aspirin as Placebo. But it is important to understand this issue, which is an example of the 'lurking variable' problem that occurs in many areas of statistics; here is a well-documented example that explains the paradox.

This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones (see the Wikipedia article on Simpson's Paradox for details). Here is a summary of the success rates of two treatments involving both small and large kidney stones.

	A	B
Small stones	81/ 87 (93%)	234/270 (87%)
Large stones	192/263 (73%)	55/ 80 (69%)
All stones	273/350 (78%)	289/350 (83%)

The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B appears more effective when considering both sizes at the same time, ignoring the subgroups.

To avoid Simpson's Paradox, a proper meta-analysis combines the estimates of risk calculated separately in each trial. But first, we need to decide what comparison of risk is appropriate to summarize the data. One possibility with the Aspirin example is a time-to-event analysis, if the main interest is in the timing of heart attacks. However, we will concentrate here just on incidence, and look at meta-analysis of time-to-event data in Section 5.6.

The risk of death in the Placebo arm varies between the trials from 8% (MRC-1) to 20% (ISIS-2): how is the treatment likely to affect this risk? We could hypothesize that treatment with aspirin has a simple effect of reducing the risk by a certain amount regardless of the underlying risk: reducing the risk by 1%, say, in all the trials. This is contentious, because risks usually combine multiplicatively rather than additively, and there are many treatments for which it would be more likely to see a greater benefit when the risk is higher. However, here the risk difference appears to be lower in the ISIS-2 trial (1.7%) than in the MRC-1 trial (2.7%), and is in fact around 2% for all except the PARIS trial, where it is –1%. However, we cannot really expect to determine the most suitable model when there are few trials, as here: we need to rely on general scientific knowledge of the area.

5.2.1. Inverse-variance method

To start with, we will accept the above hypothesis about the effect of treatment and analyse on the risk difference scale. We will also ignore here the heterogeneity due to the difference between the PARIS trial and the rest. We can then apply the inverse-variance method to the estimates of risk differences and their SEs from each study.

Table 5.2. Risk-difference estimates from the aspirin trials.

Difference	SE
–0.02770	0.01652
–0.02496	0.01307
–0.02564	0.01667
–0.02203	0.02521
–0.02314	0.01977
0.01148	0.00903
–0.01717	0.00600

The SEs here are derived from simple Normal approximations to the binomial distribution:

$$\sqrt{\{\ r_t(n_t - r_t)/n_t^3 + r_c(n_c - r_c)/n_c^3\ \}}$$

Where r_t and r_c are the numbers "responding" (dying in this example) out of n_t and n_c subjects in the Treatment and Control groups (here Aspirin and Placebo). The combined estimate (as calculated in Section 2.5) is then

–0.0134, SE 0.0042, 95% CI [–0.0216, –0.0052], p-value 0.0013

Thus, if we accept the model despite the apparent heterogeneity, we can conclude that the aspirin treatment reduces the risk of death by 1.3% on average (SE 0.4) across the mix of patients represented by these studies: that is, we expect 13 patients per thousand to be saved by the treatment, out of the average 165 per thousand who would die without it.

The calculation can be carried out with standard regression software. A simple approach is to use weighted linear regression with variance fixed at 1.0, as mentioned in Section 2.5. But it is even easier to start with the original binomial data in the table above and fit a generalized linear model using the binomial distribution and an identity link function. In effect, this is an analysis of patient-level data, though not here involving any extra covariates. It can generally be carried out for a binary response because enough is known about the individual patient responses from the summary of how many patients responded and how many did not. In the GENMOD procedure in SAS this can be done as follows, after setting up a stacked dataset called Aspirin, which should have two rows for each trial:

Study	Treatment	Deaths	Patients
MRC-1	Aspirin	49	615
MRC-1	Placebo	67	624
CDP	Aspirin	44	758
CDP	…		

Here is the SAS code to analyse this dataset.

```
proc genmod data=aspirin;
  class treatment study;
  model deaths/patients = treatment study
    /dist=bin link=id;
```

```
run;
```

In the output below, the estimate for the "Treatment Aspirin" parameter is then the combined estimate, and the Wald CI and chi-square probability correspond to the Normal statistics calculated above.

Table 5.3. Risk-difference estimates from the aspirin trials.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Limits		Chi-Square	Prob
Intercept		1	0.1214	0.0095	0.1028	0.1400	163.92	<.0001
Treatment	Aspirin	1	-0.0135	0.0042	-0.0217	-0.0053	10.39	0.0013
Treatment	Placebo	0	0.0000	0.0000	0.0000	0.0000	.	.
study	AMIS	1	-0.0112	0.0101	-0.0310	0.0087	1.22	0.2698
study	CDP	1	-0.0446	0.0112	-0.0665	-0.0227	15.91	<.0001
study	GASP	1	-0.0030	0.0155	-0.0334	0.0274	0.04	0.8463
study	ISIS-2	1	0.0767	0.0096	0.0579	0.0954	64.27	<.0001
study	MRC-1	1	-0.0216	0.0123	-0.0456	0.0025	3.09	0.0790
study	MRC-2	1	0.0205	0.0123	-0.0036	0.0447	2.77	0.0959
study	PARIS	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

The Normal approximation to the binomial used in all three of the methods for handling risk differences breaks down for risks near the extremes of the [0, 1] range. This can be handled by using better or "exact" estimators of the variance, such as that proposed by Miettinen & Nurminen (1985). The method using binomial regression can also fail, because the effects estimated for Study and Treatment in the model may extend outside the allowed range when added together in this simplistic fashion. But before embarking on more complex methods to avoid these problems, it is essential to assess whether combination on the scale of the risk difference is really appropriate. If not, the combination should be on a more appropriate scale; once this has been done, the estimate can be used to form predictions, or adjusted means, of the treatment effect on the risk

scale – as was discussed in Section 2.7.

5.2.2. Mantel-Haenszel method

The Mantel-Haenzsel method uses different weights to combine the study estimates. It was originally proposed for combining odds ratios (Mantel & Haenszel, 1959), but has been extended for risk differences (Greenland & Robins, 1985). Here is the formula for the estimate based on k studies, in which r_{ti} subjects out of n_{ti} experience an event in the treated arm in the ith study, and r_{ci} out of n_{ci} in the control, with $n_i = n_{ti} + r_{ci}$.

$$\Sigma_{i=1}^{k} (n_{ci}\, r_{ti} / n_i - n_{ti}\, r_{ci} / n_i) \,/\, \Sigma_{i=1}^{k} (n_{ti}\, n_{ci} / n_i)$$

And the s.e. of this estimate is $\sqrt{(J/K^2)}$, where

$$J = \Sigma_{i=1}^{k} (r_{ti}\,(n_{ti} - r_{ti})\, n_{ci}^3 + r_{ci}\,(n_{ci} - r_{ci})\, n_{ti}^3) \,/\, (\, n_{ti}\, n_{ci}\, n_i^2)$$
$$K = \Sigma_{i=1}^{k} (n_{ti}\, n_{ci} / n_i)$$

A confidence interval can be calculated from the s.e., and a Z test can provide a p-value. For the aspirin data, the combined estimate is

–0.0143, 95% CI [–0.0228, –0.0058], p-value 0.0010

This is not much different from the inverse-variance method here. These two methods are asymptotically equivalent, but can differ substantially when component studies are different in size and there are not very many of them.

5.3. Binary response: risk ratio

A more natural model for the way risk is changed by a treatment is multiplicative, because probabilities naturally combine in a multiplicative way. In other words, if subject i has a risk p_i of a side-effect under one treatment, it is natural to suggest he or she has a risk αp_i under the other treatment, with α constant over subjects. So the risk difference for a patient will depend on their underlying risk. This type of model can be fitted using all the methods outlined for risk differences, but working with estimates of log-relative-risk from each study. (Actual relative risks have a naturally skewed distribution, so it is best to work on the log scale.)

Table 5.4. Log relative-risk estimates from the aspirin trials.

Log rel. risk	SE
–0.2983	0.1792
–0.3577	0.1890
–0.1899	0.1239
–0.1974	0.2262
–0.1993	0.1652

0.1118	0.0880
–0.0897	0.0314

In this example, the relative risk does not look reasonably constant across studies, but we will persevere for the purpose of illustration. The SEs here are approximate ones derived using the 'delta method' (a method of deriving an approximate distribution of a function of a Normal variable):

$$\sqrt{\{ (n_t - r_t)/(r_t n_t) + (n_c - r_c)/(r_c n_c) \}}$$

Using the inverse-variance method gives the result (on the natural log scale):

–0.0902, SE 0.0275, 95% CI [–0.1441, –0.0363], p-value 0.0010

and transforming to the relative-risk scale:

0.914, 95% CI [0.866, 0.964]

This can be interpreted as an 8.6% relative reduction in risk due to treatment (1–0.914 = 0.086), though such a statement is liable to be misinterpreted because the base of the percentage is itself a proportion or percentage. An individual patient who would have 16.5% risk under Placebo is expected on average to have a 15.1% risk (0.165×0.914 = 0.151), which is a 1.4% absolute reduction in risk.

These results can be derived either by explicitly calculating the weighted estimate or by using weighted regression, as for risk difference. They can also be derived using a patient-level approach, fitting a generalized linear model with a log link (rather than the identity link needed for risk difference). Again, the g.l.m. can break down with some data, when the risks are large and the multiplicative model conflicts with the upper limit of 1.0 for risks.

5.3.1. Mantel-Haenszel method

The Mantel-Haenszel method may also be used for relative risks, as described by Rothman & Greenland (1998). The formula for the estimator is

$$\Sigma_{i=1}^{k} (n_{ci}\, r_{ti}/n_i) \;/\; \Sigma_{i=1}^{k} (n_{ti}\, r_{ci}/n_i)$$

With s.e. (on the log scale) $\sqrt{(P / RS)}$, where

$$P = \Sigma_{i=1}^{k} ((r_{ti} + r_{ci})\, n_{ti}\, n_{ci} - r_{ti}\, r_{ci}\, n_i) / n_i^2$$
$$R = \Sigma_{i=1}^{k} (r_{ti}\, n_{ci} / n_i),\; R = \Sigma_{i=1}^{k} (r_{ci}\, n_{ti} / n_i)$$

This allows derivation of a confidence interval and p-value:

0.914, 95% CI [0.866, 0.964], p-value 0.0010,

which is virtually identical to the result from the inverse-variance method.

5.4. Binary response: odds ratio

Meta-analysis of binary outcomes is most commonly done with odds ratios. This is because the odds ratio is the natural way to compare probabilities, taking into account their range, [0, 1], and the fact that probabilities in the centre of the range are inherently more variable than ones at either edge. For small probabilities, there is very little difference between odds and risk ratios; for example, if the risk under one treatment is 0.01 and the relative risk is 2, the odds ratio is 2.02; for 0.02 and 2 it is 2.04; and even for 0.05 and 2 it is 2.11 (only just 5% different from the relative risk). But the odds ratio scale works for large probabilities near 1.0 as well, while risk ratios do not make much sense: again, it is better to work on the log-odds-ratio scale.

Table 5.5. Log odds-ratio estimates from the aspirin trials.

Log odds-ratio	SE
–0.3289	0.1972
–0.3845	0.2029
–0.2196	0.1431
–0.2222	0.2545
–0.2255	0.1876
0.1246	0.0981
–0.1110	0.0388

Like the relative risks, the odds ratios do not appear to be constant across these studies, but we will again persevere for illustration. The SEs are obtained using the Normal approximation to the Poisson distribution, which gives the approximate formula:

$$\sqrt{\{\ 1/r_t + 1/(n_t - r_t) + 1/r_c + 1/(n_c - r_c)\ \}}$$

Using the inverse-variance method gives the result (on the natural log scale):

–0.1088, SE 0.0331, 95% CI [–0.1737, –0.0438], p-value 0.0010

and transforming to the odds-ratio scale:

0.897, 95% CI [0.841, 0.957]

This can be interpreted as a 10% reduction in odds due to treatment (1–0.897=0.10, to two significant figures). An individual patient with 16.5% risk (=19.8% odds) under Placebo is expected on average to have odds of 17.7% (0.198×0.897=0.177), which is the same as 15.1% risk (0.177/(1+0.177)=0.151), equal to the summary from the analysis with relative risks.

As before, these results can be derived either by explicitly calculating the weighted estimate, using constrained regression, or fitting a generalized linear model – this time with the canonical logit link function; i.e., logistic regression.

```
proc genmod data=aspirin;
  class treatment study;
  model death/patients = treatment study /dist=bin;
```

```
run;
```

Whereas the g.l.m. method can fail for risk differences and ratios, it cannot for odds ratios because combination of the estimates of the Study and Treatment effects cannot extend outside the allowable range on this scale. The Normal approximation is also more robust when applied to small probabilities analysed on the logit scale.

5.4.1. Scoring method

The method above is based on likelihood, and the estimate is the maximum-likelihood estimator given the assumptions. A simpler approach to estimation has been widely used in meta-analysis, using the idea of 'efficient scores' and Fisher's information. This can be seen as a one-step approximation to the likelihood method; but the approximation is often good (unless the data are extreme with respect to the distributional assumptions) and the calculations are certainly easier – though the availability of computers now removes the original force of this argument. The scoring method also has some beneficial side-effects, as we shall see in the next section. The estimates of the log-odds ratio and SE are as follows.

Table 5.6. Log one-step odds-ratio estimates from the aspirin trials.

Study	logOR	SE
MRC-1	–0.3264	0.1951
CDP	–0.3803	0.1996
MRC-2	–0.2188	0.1425
GASP	–0.2218	0.2537
PARIS	–0.2315	0.1923
AMIS	0.1245	0.0979
ISIS-2	–0.1109	0.0388

These are very close to the estimates from logistic regression, as is to be expected when the treatment effect is small. The values can be calculated with the following formulae.

$\text{logOR} = (o_i - e_i)/v_i$

$\text{SE(logOR)} = \sqrt{\{1/v_i\}}$

where o_i can be thought of as the observed number of events in the Treatment arm ($=r_{ti}$), and e_i as the expected number, calculated as

$(r_{ti} + r_{ci})n_{ti} / n_i$

and v_i is calculated as:

$(r_{ti} + r_{ci})(n_i - (r_{ti} + r_{ci}))\, n_{ti}\, n_{ci}/n_i^3$

Combining these with the usual inverse-variance method gives the following results (on the natural log scale):

–0.1089, SE 0.0331, 95% CI [–0.1738, –0.0440], p-value 0.0010

and transforming to the odds-ratio scale,

0.897, 95% CI [0.841, 0.957]
which is virtually the same as from the likelihood approach.

5.4.2. Conditional logistic method

Another way of analysing binary data is to condition on the observed margins: in this case, the total number of events in each study. This margin is a measure of the average event rate in each study, and can be seen to be ancillary to the effect of interest – the way the rate differs between treatments within each study. Rather than using the binomial distribution to model the observed counts, the hypergeometric distribution is appropriate when the margin is treated as fixed.

Using this method on the aspirin trial gives the following result (virtually identical here to the unconditional analysis):
–0.1089, SE 0.0331, p-value 0.001
and transforming to the odds-ratio scale:
0.897, 95% CI [0.840, 0.957]
This can be done with SAS using the LOGISTIC procedure, setting the STRATA statement to specify conditioning with respect to the Study classification:

```
proc logistic data=aspirin;
  class treatment (param=ref) study;
  model deaths/patients = treatment;
  strata study;
  run;
```

It is helpful to avoid LOGISTIC's default 'effect' parameterization by setting the PARAM option in the CLASS statement, to give the usual estimate of treatment difference.

5.4.3. Peto method

A simpler version of the conditional approach is provided by a method put forward by Yusuf, Peto, Lewis, Collins & Sleight (1985). This uses the scoring method, and is virtually the same as the scoring method for the unconditional approach: the only difference is that the formula for the variance v_i is calculated as:
$(r_{ti} + r_{ci})(n_i - (r_{ti} + r_{ci}))\, n_{ti}\, n_{ci} /(n_i^2(n_i-1))$
which is a factor $n_i/(n_i-1)$ times the previous formula.

The Peto method is not recommended for general use. The estimates of the odds-ratio that it uses are clearly biased if there is a substantial treatment effect. (For example, the OR of a 10% risk to a 20% risk is 0.444, while the Peto OR is 0.458 – a 3% bias; for 10% to 30% the OR and Peto OR are 0.259 and 0.288 – an 11% bias.) It has also been shown to be biased when the data are very unbalanced (with a ratio of subjects on the two treatments of the order of 8:1 or more). However, it is generally accepted as the best

method to use when events are rare, the treatment effect is not large, and data are reasonably balanced. This is primarily because it allows inclusion of studies that have no events in one or other treatment arm (though it uses no information from studies with no events at all, in common with all the other non-Bayesian methods).

5.4.4. Mantel-Haenszel method

The Mantel-Haenszel method was originally designed for odds ratios (Mantel & Haenszel, 1959). The formula for the estimator is:

$$\Sigma_{i=1}^{k}((n_{ci} - r_{ci})r_{ti}/n_i) \,/\, \Sigma_{i=1}^{k}((n_{ti} - r_{ti})r_{ci}/n_i)$$

The s.e. of this estimator (on the log scale) is:

$$\sqrt{[\ 0.5\sum\{\ A_iC_i/C^2 + (A_iD_i+B_iC_i)/CD + B_iD_i/D^2\}]}$$

where

$$A_i = (r_{ti} + n_{ci} - r_{ci})/n_i\ ,\ \ B_i = (r_{ci} + n_{ti} - r_{ti})/n_i\ ,$$
$$C_i = r_{ti}\ (n_{ci} - r_{ci})/n_i\ ,\ \ D_i = r_{ci}\ (n_{ti} - r_{ti})/n_i\ ,$$
$$C = \sum C_i^2\ ,\ D = \sum D_i^2$$

This gives the result (on the natural log scale):

–0.1088, SE 0.0341, 95% CI (–0.1756, –0.0421) p-value 0.002

Transforming to the odds-ratio scale:

0.897, 95% CI (0.841, 0.957)

This method can be carried out using the FREQ procedure in SAS:

```
proc freq data=aspirin2;
  tables study*treatment*response /cmh2;
  run;
```

though you need to construct the dataset Aspirin2 to have a row for every subject, with a column (Response) storing ‘yes’ or ‘no’ according to whether or not the subject died.

5.4.5. Exact methods

There are some methods that do not rely on the asymptotic approximations of the methods above. They are often called “exact” methods, though this name should not be interpreted as an assertion of superiority over the asymptotic methods: just that the treatment of variability is handled in a more precise way. Exact methods still depend on assumptions, and can sometimes be unnecessarily conservative. The LOGISTIC procedure in SAS has an EXACT statement, which can provide some results. We do not describe these methods here, but just refer to a review by Emerson (1994).

5.5. Other types of response

5.5.1. Poisson and negative-binomial response

When the response variable of interest is a count, such as number of exacerbations in asthma, for example, the usual analysis is Poisson regression using a log scale for combining effects of treatment and covariates. Alternatively, if the Poisson assumption (that the process generating events has no "memory") is not reasonable, and the counts show substantial heterogeneity compared to Poisson data, negative binomial regression may be used. The results from such trials can be combined using the inverse-variance method as above, working with estimates from each study of the average count under each treatment, or the average rate if the analysis incorporates information about varying exposure.

A patient-level analysis can be done using Poisson regression of the total counts in each treatment group of each study, in the same way as the logistic regression of binary data analysed on the log-odds-ratio scale above. The Poisson regression can be adjusted using an offset variable to take account of total exposure to the treatment in each group. But negative binomial regression requires the individual patient counts, because of the extra dispersion parameter that is involved. Of course, further covariates can be included in the model as for a continuous response, if the covariate information is available. Here is the minimal SAS code needed for patient-level meta-analysis of Poisson data.

```
proc genmod data=asthma;
  class treatment study;
  model exac=study treatment /dist=poisson link=log;
  run;
```

5.5.2. Ordinal response

Variables such as Clinical Global Improvement are recorded on an ordinal scale, with categories such as "No change", "Slight improvement", "Improvement", and so on. Whereas these are sometimes analysed simply by combining categories to derive a binary response (e.g. "Improve" vs "Not improve"), the full information can be used in a multinomial model. The most commonly used is the proportional-odds model, which is analogous to the odds model for binomial data with an assumption that treatment and covariate effects have a consistent effect in terms of odds calculated at any "cut-point": i.e. combining categories above and below any point on the ordinal scale.

Information on ordinal response variables can be combined using the inverse-variance method as usual. A patient-level analysis can be achieved by including a Study effect to give a stratified proportional-odds model. This can be fitted using the GENMOD or NLMIXED procedures in SAS, as illustrated in Sections 5.4.7 and 5.4.2 of Whitehead (2002).

```
proc genmod data=ordinal;
```

```
class treatment study;
model cgi=study treatment
  /dist=multinomial link=cumlogit;
run;
```

5.5.3. Time-to-event response

Time-to-event studies are often modelled with the proportional-odds, or Cox model. The estimates of log hazard ratio from this model fitted to each study can be combined with the inverse-variance method. A common alternative is the log-rank statistic. For a patient-level analysis, Study is introduced into the proportional-odds model as a stratifying covariate, as shown in the following code for SAS using the PHREG procedure.

```
proc phreg data=survival;
  model time*censor(0)=treatment /ties=discrete;
  strata study;
  run;
```

The variable 'censor' is a 0/1 variable recording whether or not a subject's observation was censored at the end of the study, with '(0)' indicating that the value 0 represents censoring. Again, this is illustrated in Whitehead (2002), in Section 5.5.2.

A review by Smith & Williamson (2007) indicated that there is little difference in practice between meta-analysis of estimates from the Cox model or from log-rank analysis. They appear to have a similar relationship to that between the likelihood and scoring methods for binary data. The review also showed that individual patient meta-analysis using a stratified Cox model also gave similar results to the corresponding summary-level meta-analysis.

References

Emerson JD (1994). Combining estimates of the odds ratio: the state of the art. *Statistical Methods in Medical Research* **3**:157–178.

Fleiss JL, Gross AJ (1991). Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* **44**:127–139.

Greenland S, Robins JM (1985). Estimation of common effect parameter from sparse follow up data. *Biometrics* **41**:55–68.

Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute* **22**:719–748.

Miettinen OS, Nurminen M (1985). Comparative analysis of two rates. *Statistics in Medicine* **4**:213–226.

Rothman KJ, Greenland S. (1998) *Modern Epidemiology* (2nd edition). Philadelphia: Lippincott-Raven.

Smith CT, Williamson PR (2007). A comparison of methods for fixed-effect s meta-analysis of individual patient data with time to event outcomes. *Clinical Trials* **4**:621–630.

Whitehead A (2002). *Meta-analysis of Controlled Clinical Trials*. Chichester: Wiley.

Yusuf S, Peto R, Lewis J, Collins R, Sleight P (1985). Beta-blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* **27**:335–371.

6. RANDOM-EFFECTS APPROACHES

6.1. The DerSimonian-Laird method

Random-effects models incorporate variation among studies into the estimate of the combined effect measure. This may have only a slight effect on the estimate itself, but often causes a substantial increase in the SE of the estimate. The most commonly used method, proposed by DerSimonian and Laird (1986), is a variation of the inverse-variance method (see Section 2.5): it adjusts the weight given to each study to incorporate heterogeneity across studies. Alternative estimates are derived by using simple or profile likelihood methods (see Section 6.4).

The DerSimonian-Laird method decomposes the observed variance into two component parts, within-studies and among-studies, and then use both parts when assigning a weight to each study. The mechanism used to decompose the variance is to estimate the total variance and the within-studies variance. The difference between these two values will give us an estimate of the between-studies variance, denoted by τ^2.

The *Q* statistic represents the total variance relative to the within-study variance, and is defined as

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}_\bullet\right)^2$$

where $\hat{\theta}_i$ = estimate of the effect in the *i*th study, w_i = weight given to the *i*th study,

$$\hat{\theta}_\bullet = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} = \text{ combined estimate from the FE analysis}$$

and k = the number of studies. The value of w_i is the same as the weight used in the fixed-effects analysis, i.e. the variance of $\hat{\theta}_i$.

Q has a χ^2 distribution under the null hypothesis of no heterogeneity, with $k - 1$ d.f., so

$$\mathrm{E}(Q) = k - 1 .$$

This allows us to estimate the between-studies variance, τ^2, as

$$\tau^2 = \begin{cases} \dfrac{Q - (k-1)}{C} & \text{if } Q > k-1 \\ 0 & \text{if } Q \le k-1 \end{cases}$$

where

$$C = \sum w_i - \frac{\sum w_i^2}{\sum w_i}$$

The numerator $Q - (k - 1)$ is a measure of the excess (observed minus expected) variance. The denominator, *C*, is a scaling factor that has to do with the fact that *Q* is a weighted sum of squares. It ensures that τ^2 is on the same scale as the within-studies variance.

These calculations can be applied to the aspirin data used in the FE analysis in Section 5.2.1. The value of Q is

$$Q = 20.7762 - \frac{(-770.2181)^2}{57326.9} = 10.42787 .$$

This leads to C,

$$C = 57326.9 - \frac{992123812}{57326.9} = 40020.5$$

and then τ^2

$$\tau^2 = \frac{10.42787 - (7-1)}{40020.5} = 0.0001106 .$$

This is our estimate of the variance in the true effect of aspirin among studies.

Assigning weights under the random-effects model

In the random-effect analysis, as in the fixed-effects analysis, each study is weighted by the inverse of its variance, but this variance now includes not only the original (within-studies) variance but also the between-studies variance, τ^2.

Thus under the random-effects model the variance assigned to each study is

$$V_i^* = V_i + \tau^2$$

where V_i is the variance of the FE estimate as before, and the weight assigned to each study is

$$w_i^* = \frac{1}{V_i^*}.$$

The calculation of the random-effects combined estimate then follows the same pattern as that of the fixed-effects combined estimate in (6.2), namely

$$\hat{\theta}_\bullet^* = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*}.$$

This is an estimate of $\theta_\bullet^*$, the mean of the probability distribution in the upper part of Fig. 6.1. The variance of this combined estimate is given by the reciprocal of the sum of the weights, that is

$$V^* = \frac{1}{\sum_{i=1}^{k} w_i^*},$$

and its standard error is the square root of the variance, that is,

$$SE\left(\hat{\theta}_\bullet^*\right) = \sqrt{V^*}.$$

These methods can be applied to the aspirin data, giving the random-effects combined estimate

$$\hat{\theta}_\bullet^* = \frac{-360.164}{24105.96} = -0.0149,$$

its variance

$$V^* = \frac{1}{24105.957} = 0.00004148,$$

and standard error 0.00644.

6.2. The DerSimonian-Laird method using SAS

The RE estimate and standard error can also be obtained by a weighted least-squares regression analysis, in which the observed responses (y) are the study estimates of

treatment difference $\hat{\theta}_i$, the weights are the random-effect weights w_i^*, and there are no explanatory variables in the model, only a constant term. The method is available in SAS PROC MIXED. The data are arranged for analysis by SAS in the following spreadsheet:

study	est_rd	weight
MRC-1	-0.027697	2610.447
CDP	-0.024962	3554.921
MRC-2	-0.025639	2576.019
GASP	-0.022031	1343.698
PARIS	-0.023141	1996.167
AMIS	0.011482	5205.503
ISIS-2	-0.017165	6819.203

After this spreadsheet is loaded into a SAS dataset called 'weights', the following code produces the analysis required:

```
proc mixed data=weights noprofile;
  model est_rd = /solution;
  weight weight;
  parms (1) /noiter;
run;
```

The results include the following:

Solution for Random Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-0.01494	0.006441	6	-2.32	0.0595

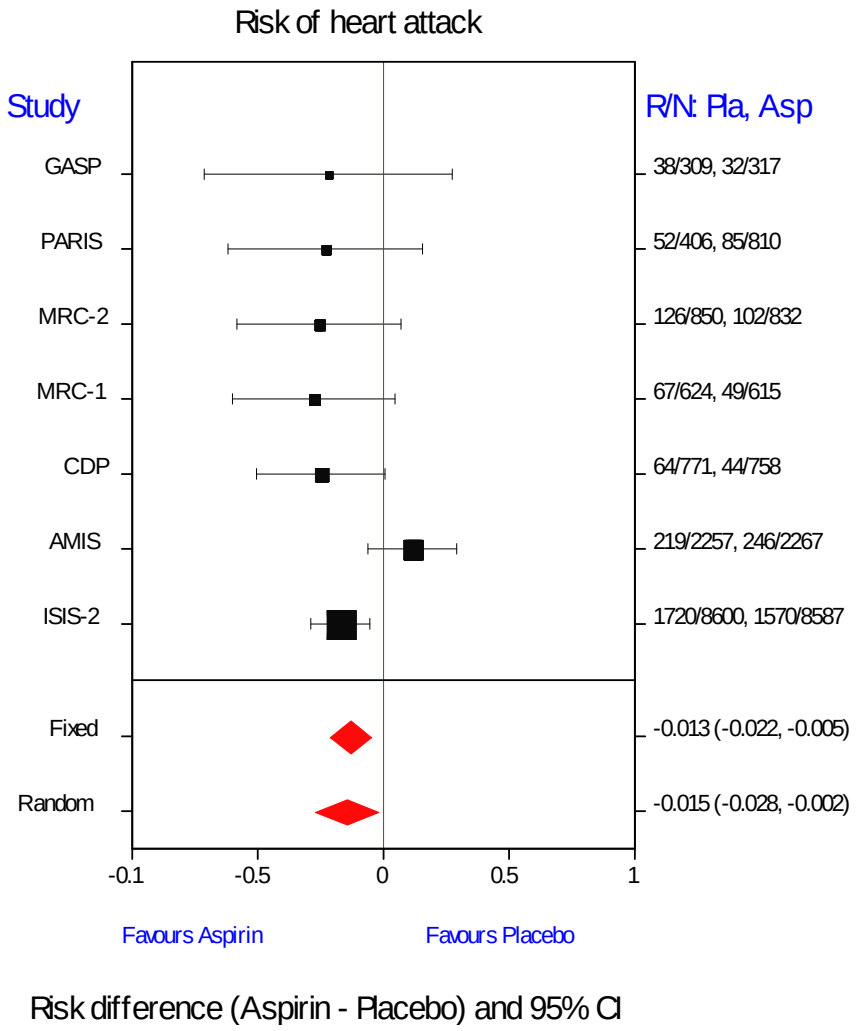
6.3. Impact of the RE model on the estimate and confidence interval

The results of applying the RE model to the Aspirin data set are illustrated in a forest plot in Fig. 6.3. The results of the FE and RE meta-analyses are compared by the two diamonds at the bottom of the figure. The main difference is that the RE combined estimate is less precise than the FE value (FE analysis: estimate = −0.01348, SE = 0.004177; RE analysis: estimate = −0.01494, SE = 0.006441). It is important to remember that this is not a weakness: it reflects the fact that the FE value represents only the studies used in the analysis, whereas the RE value is representative of the population from which the studies are drawn. Under an FE model, we implicitly set the between-studies dispersion to zero. Therefore, for the purpose of estimating the mean effect, the only source of uncertainty is within-study error. With random effects, dispersion between

studies is considered a real source of uncertainty, is estimated from the data, and is included in the standard error of the combined estimate. The effect estimates from these six studies vary more than would often occur by chance. If this variation is real, it means that the effect of aspirin varies depending on factors, known or unknown, that vary from study to study. For example, if one study used older subjects than another, or a shorter duration of observation, the effect size might be different.

Note also that the RE combined estimate is slightly further from zero (more negative) than the FE value. This is because of the difference in the impact of the AMIS study, the only study that gave a positive estimate. The forest plot shows that this was a large study with high precision, which therefore made a strong contribution to the FE combined estimate. However, its influence on the RE value is diluted by τ^2, which contributes equally to the variance assigned to each study (Equation 6.10).

Figure 6.3. Forest plot of Aspirin meta-analysis.



6.4. Measuring heterogeneity

The Q Statistic

The classical measure of heterogeneity is Cochran's *Q*, which is calculated as above. *Q* has low power as a test of heterogeneity when the number of studies is small, as it is in most meta-analyses (Gavaghan et al., 2000), but when the number of studies is large the power of *Q* is great (Higgins et al, 2003), and it may give a statistically significant result even if the amount of heterogeneity is too small to be of clinical significance. Fleiss suggested compensating for low power of *Q* using a 0.10 cut-off rather than the conventional 0.05. However, this is arbitrary, and will not provide a strong defence against a false-negative result when the power is very low.

Although the results of statistical tests for heterogeneity provide useful descriptive information about variability between trials, a decision based purely on the *p*-value is not to be recommended. The heterogeneity test results should be considered alongside a qualitative assessment of the combinability of studies in a systematic review. Moreover, it is necessary to distinguish between a statistically significant difference and a clinically important difference. Hardy and Thompson (1998) concluded that the power of the heterogeneity test can be low, and its value limited, in the case of sparse data (when the total information, the sum of the weights, is low) or when there is large variability among the weights of the trials, especially if one trial has a much larger weight than the others.

Design flaws in primary studies, and publication bias, may compromise the interpretation of the *Q* statistic. Design flaws lead to estimates of different effects compared to well-designed ones, so heterogeneity is inflated. Conversely, publication bias may cause the studies that are available for meta-analysis to be less heterogeneous than the full range of relevant studies.

The I^2 Statistic

The I^2 statistic measures the percentage of variation across studies that is due to heterogeneity rather than chance (Higgins and Thompson, 2002; Higgins et al., 2003). It is calculated as

$$I^2 = 100\% \times (Q - (k-1))/Q \qquad (7.1)$$

I^2 is a simple and intuitive expression of the lack of consistency among the results of different studies. Unlike Q it does not inherently depend upon the number of studies considered. A reasonable interpretation of the I^2 value is as follows:

- $I^2 < 50\%$: low heterogeneity among studies
- $50\% < I^2 < 75\%$: intermediate heterogeneity
- $I^2 > 75\%$: high heterogeneity.

A confidence interval for I^2 can be constructed using either the iterative non-central chi-squared distribution method of Hedges and Piggott (2001) or the test-based method of Higgins and Thompson (2002).

In the case of the aspirin data (section 6.2),

$$I^2 = 100\% \times (10.42787 - (7-1))/10.42787 = 42\%,$$

indicating that nearly half the observed variation in the effect estimates among trials is due to real heterogeneity.

The L'Abbé plot (Section 4) can be used to explore the lack of consistency among studies visually.

6.5. The likelihood approach to random effects (REML/ML)

A random-effects meta-analysis can also be performed using the maximum-likelihood (ML) or restricted maximum-likelihood (REML) approach, implemented for example in SAS PROC MIXED. This method uses as its starting point the individual study estimates and their fixed-effect weights, rather than the random-effect weights used above. Hence the calculation of τ^2 is not required as a preliminary step: this variance component is estimated as part of the fitting process. The variances of the individual study estimates, and hence their fixed-effect weights, are assumed to be known without error.

The data required for this approach are arranged for analysis by SAS in the spreadsheet below:

study	est_rd	feweight
MRC-1	-0.027697	3670.639
CDP	-0.024962	5859.738
MRC-2	-0.025639	3602.931
GASP	-0.022031	1578.355
PARIS	-0.023141	2562.027
AMIS	0.011482	12275.918
ISIS-2	-0.017165	27777.315

When this spreadsheet is located in the sheet ‘feweights’ in the Excel workbook ‘ISIS data (local).xls’, the following code produces the analysis required:

```
PROC IMPORT OUT= WORK.FEWEIGHTS
            DATAFILE= "&rootdir.ISIS Data (local).xls"
            DBMS=EXCEL REPLACE;
     SHEET="'feweights$'";
     GETNAMES=YES;
     MIXED=YES;
```

```
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

data mafeweights;
  set feweights;
  var = 1/feweight;
  /*Specify the G matrix such the element (row, col)
  holds the corresponding value of var.
  All other elements are set to 0*/
  col = _n_; row = _n_; value = var;
run;

ods rtf file="&rootdir.aspirin2.rtf";
proc mixed data=mafeweights order=data method=ml;
   /*Can also set method = reml*/
   Class study;
   Model est_rd = /solution ddfm=kr;/*model is intercept only,
      print predicted and solution for FE;
      use KR option, because there are usually few df for var*/
   Random study/gdata=mafeweights;/* G matrix in dataset 'maweights'*/
   Repeated diag; /* Sampling Variance 'R' is Diagonal*/
run;
ods rtf close;
```

The output produced by PROC MIXED includes the following:

Convergence criteria met.

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.000099

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-0.01479	0.006669	6.09	-2.22	0.0677

The values obtained are similar to those obtained previously, but not identical. The estimate of the covariance parameter 'Residual', 0.000099, is the value of τ^2. Again, it is similar to that obtained earlier, but not identical.

In order to fit the model using the restricted maximum-likelihood criterion, instead of the ordinary maximum-likelihood criterion, the option setting 'method=ml' is changed to 'method=reml' in the SAS code. The corresponding part of the output produced by PROC MIXED is then as follows:

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.000138

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-0.01523	0.007165	5.98	-2.13	0.0778

The estimates obtained are somewhat changed by the change of method.

The DerSimonian and Laird method does not adequately reflect the error associated with parameter estimation, especially when the number of studies is small. The profile-likelihood method usually provides an estimate with better coverage probability and should be used when possible (Brockwell & Gordon, 2001).

For odds ratios, a logistic random-effects model could be used to combine results, although it may underestimate the uncertainty (Smith, Spiegelhalter et al., 1995).

6.6. RE models for individual patient meta-analysis

If individual-patient data are available from the studies to be included in a meta-analysis, the random-effects model can be applied directly to these. The principles and methods are the same as in the analysis of multi-centre trials, with one exception. The analysis uses a mixed model, because it contains both fixed effects (the main effect of each study and the main effect of the treatment) and random effects in addition to the residual effects (the study × treatment interaction effects). The exception is that it is unusual to specify the main effects of studies as fixed when their interaction is specified as random.

If the study × treatment interaction effects are specified as random, they are implicitly regarded as a random sample from an infinite population of such effects. As there are only two treatments, placebo and active, this implies the existence of an infinite

population of potential or actual studies, from which the studies in the meta-analysis are similarly sampled. Therefore the main effects of these studies are also randomly sampled from an infinite population, and it is strictly speaking contradictory to specify the interaction effects as random but the main effects of study as fixed. Nevertheless, there is a powerful argument for doing so: if study effects were specified as random, between-study variation will contribute to the estimate of the treatment effect. This would introduce a component of the treatment estimate that is not protected by the randomization within each contributing trial. One way to exclude this is to use an implementation of the maximum-likelihood method that allows the between-trials component to be excluded: the study effects could then also be specified as random. The alternative is to use the device of specifying the study effects as fixed, which in fact leads to the same results as the first way.

References

Brockwell SE, Gordon IR (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**:825–840.

DerSimonian R, Laird N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**:177–88.

Gavaghan DJ, Moore RA, McQuay HJ (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain* **85**:415–424.

Hardy RJ, Thompson SG (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* **17**:841–856.

Hedges LV, Pigott TD (2001). The power of statistical tests in meta-analysis. *Psychological Methods* **6**(3):203–217.

Higgins JP, Thompson SG (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**:1539–1558.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* **327**(7414):557–560.

Smith TC, Spiegelhalter DJ, Thomas A (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**:2685–2699.

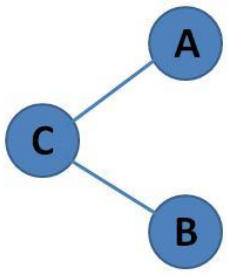
7. BRIEF INTRODUCTION TO NETWORK META-ANALYSIS AND BAYESIAN METHODS

All the techniques of meta-analysis we have looked at so far have been developed to compare two treatments, based on studies that tested both of them. Of course, some of the studies may have included other treatments as well. In a summary-level meta-analysis, the summaries from the other treatments will not be included in the meta-analysis; however, the presence of the other treatments in the studies may have had some effect on the summaries (particularly the standard errors) of the treatments that are to be compared. In a patient-level meta-analysis, usually only data from the patients who were given the two treatments will be included, so information from the other treatments will be ignored.

7.1. Indirect comparison

When there is not enough evidence from trials that compare the two treatments directly, it is possible to include evidence from indirect comparisons. For example, there may be a set of studies comparing one treatment to control, another set comparing the other treatment to the same control, and perhaps a third set comparing the two treatments (maybe also with control). Many approaches have been put forward for making use of the indirect comparisons. For summary-level data with non-overlapping information (i.e. no direct comparisons), the simplest approach is to extend one of the standard two-stage approaches (forming summaries, then combining them) to add a third stage. This stage simply uses the combined estimates comparing each treatment against a common control, and their variances, to construct an estimate comparing the two treatments themselves.

Figure 7.1. Indirect MA



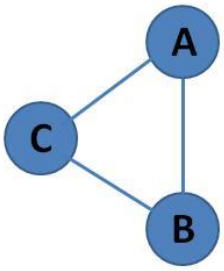
For example, consider the comparison of two drugs A and B, and a control (C), with information available as in Figure 7.1. Say the combined fixed-effect estimate of the difference between Treatments A and C (control) is d_{AC} with s.e. s_{AC}, and for Treatments B and C it is d_{BC} with s.e. s_{BC}. Then an estimate of the difference between Treatments A and B is $d_{AC}–d_{BC}$, with s.e. $\sqrt{\{s_{AC}^2+ s_{BC}^2\}}$ because d_{AC} and d_{BC} are independent.

This method is called adjusted indirect comparison, and was described in detail by Bucher et al. (1997). It is clear that evidence from head-to-head comparisons is much to be preferred, because in practice there will always be room for doubt as to whether two sets of trials are fully comparable. Indirect comparison is also far less efficient in terms of numbers of subjects whose results need to be incorporated in the analysis. In the idealized situation of all trials being exactly the same in terms of size, treatment effect and so on, four times as many trials are needed to give the same power as in a direct comparison. For one trial, suppose that the estimated treatment effect has standard error s. Then for a

meta-analysis of m trials, all of the same size and assuming a common true treatment effect, an inverse-variance meta-analysis would provide an s.e. of the treatment effect of $s\sqrt{\{1/m\}}$. The s.e. from an indirect comparison based on $n/2$ trials for each comparison (so n trials in all) is $\sqrt{\{s_{AC}^2 + s_{BC}^2\}} = \sqrt{\{s^2/(n/2) + s^2/(n/2)\}} = s\sqrt{(4/n)}$. Hence, n needs to be four times m to give the same expected s.e.

The adjusted indirect comparison method outlined above is a special case of multi-treatment meta-analysis models (Gleser & Olkin, 1994). It can be seen as a simple kind of meta-regression. The method can be carried out on summary-level data by analysing the treatment effects in a simple meta-regression model with a covariate indicating which treatment comparison is supplied from each trial. The estimate is then constructed as a contrast from the fitted model, in the standard way, as illustrated in the example below. With patient-level data, all that is needed is a covariate indicating which treatment the patient received. The analysis relies on the invariance of the treatment effects across study populations and retains the benefits of randomization in the original RCTs.

Figure 7.2. MTC

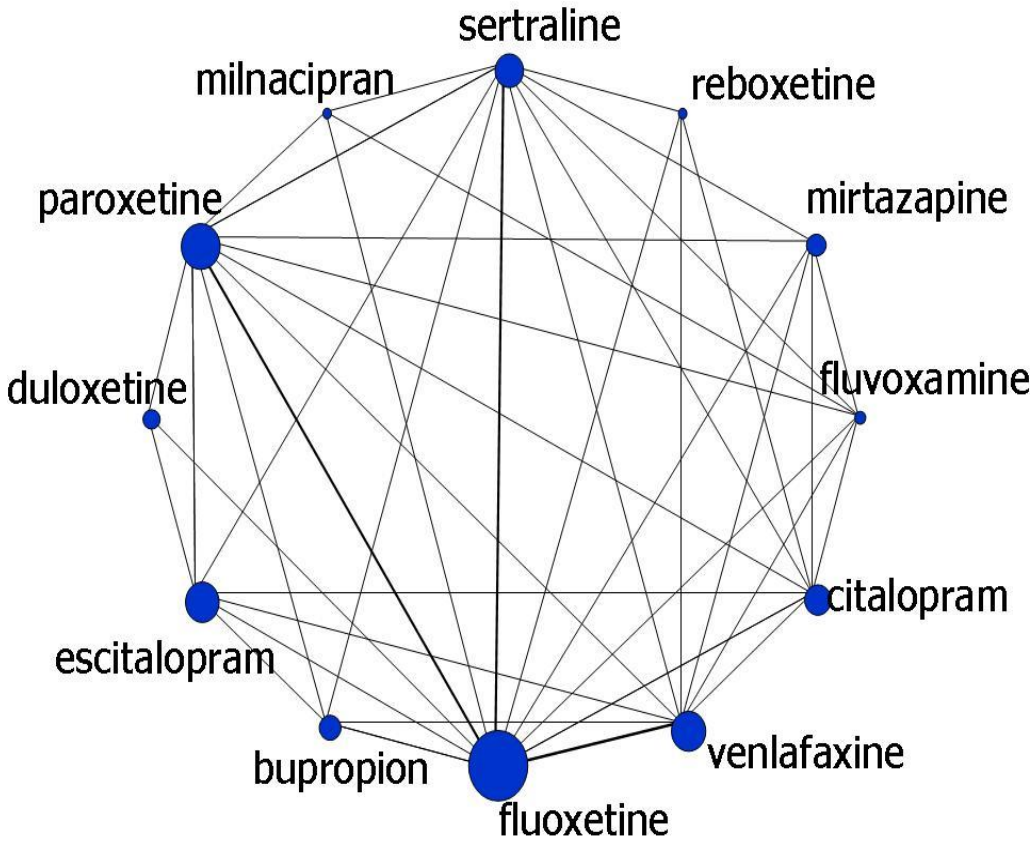


The simplest extension of this method for simple indirect comparison is when direct comparisons are also available, as shown schematically in Figure 7.2. It is then called a "Mixed Treatment Comparison" (MTC) because the method mixes information from direct and indirect comparisons. The same approach can also be used for more complicated situations involving chains of comparison, such as A vs. B, B vs. C and C vs. D, as we will see in Section 8.3.

7.2. Network meta-analysis

The term "Network meta-analysis" is applied to any analysis that makes use of more than just direct comparisons. It would be pretentious to use the term for a simple adjusted indirect comparison involving just two drugs compared with a control, shown in Figure 7.1, or the MTC shown in Figure 7.2, though they can be seen as simple examples. It is more generally used to describe the comparison of a set of treatments using data from trials that compare subsets of them. This is the backbone of comparative effectiveness research (CER). Figure 7.3 shows a schematic summary of the network formed by 19 meta-analyses of pairs of antidepressant drugs that were published over a two-year period, involving 12 drugs in all. Some drugs were not compared directly in any of these meta-analyses, but the network shows how indirect comparisons can be made using the comparisons that are available.

Figure 7.3. Network of 12 antidepressants



For example, if we wished to compare paroxetine with fluoxetine, there is some direct information available from head-to-head trials summarized in published meta-analyses. But in addition, there is a comparison of paroxetine with fluvoxamine, and another of fluvoxamine with fluoxetine: this provides indirect information to be added into an MTC of paroxetine and fluoxetine. This can also be done with all the other drugs except reboxetine, which has not been compared with paroxetine directly. But reboxetine has been compared with sertralin, for example, so there is a three-link chain paroxetine–sertralin–reboxetine–fluoxetine. A full network analysis uses all the available chains, and will typically produce estimates of all the pairwise comparisons.

Network meta-analyses can be carried out using precisely the same method as illustrated in Section 8.2. The only difference is that there are more treatments, i.e. more potential values for the Treat column in the dataset. If the outcome is binary, logistic regression can produce the maximum-likelihood estimates of all possible pairwise comparisons. The same approach can be used with linear models for continuous outcomes, Poisson and negative binomial models for counted outcomes, and Cox models for time-to-event outcomes.

An indirect comparison is often considered to be effectively an observational study. For example, the Cochrane Handbook for systematic reviews of interventions 4.2.4 (Cochrane Library Issue 2) says:

> "Indirect comparisons are observational studies across trials, and may suffer the biases of observational studies, for example confounding."

Various authors (including Victor, Egger and Moher) have each claimed that pair-wise MA is also "observational", even when stratifying by study and relying only on randomized clinical trials. The reasons given are that MA is usually retrospective, does not use a random sample of studies, and that an MA itself is not a randomized study. For more details, see Lu & Ades (2006), Song et al (2008), Song et al (2009), and Ades (2009).

Differences in factors between the two sets of trials that could influence the outcome will bias any indirect comparison. In addition, of course, all the problems with any other meta-analysis, such as heterogeneity, publication bias, or differing control event rates, also apply for indirect comparisons.

In general, indirect comparisons have low power and often lead to indeterminate results. There is no consensus on how to interpret results that differ substantially from direct evidence or on how to weigh findings of indirect comparisons against those results from non-randomized direct evidence. A guide to indirect and network methods has been drafted by ISPOR (International Society for Pharmacoeconomics and Outcomes Research) which gives more information and references (Janssen et al., 2011). A PSI working group published an introductory guide to these methods (Jones et al., 2011).

7.3. Bayesian methods

Recently, there has been a lot of interest and application of Bayesian methods for network analysis, and this has become the de facto standard for CER. In fact, the estimates of treatment comparisons produced by Bayesian methods should generally be almost identical to those produced by frequentist methods, because there is no prior information available. Any differences between the two approaches are likely to be caused by the use of prior distributions that are inadequately non-informative, or by the expected simulation error that is inherent in the Monte Carlo Markov-Chain (MCMC) method. This is also the case for SEs and CIs of the estimates from fixed-effect MA. However, there can be differences for random-effects MA of non-Normal data because Bayesian methods take account of the uncertainty of the random-effects variance itself, while frequentist methods do not.

The popularity of the Bayesian approach is the appeal of being able to make probability statements about the estimates, such as the probability of a given treatment being the best among a set that have been compared, and the ability to use the simulated posterior

distribution to study functions of parameters. The Bayesian approach to network meta-analysis is also better-developed and more publicized.

The WinBUGS software provides for Bayesian meta-analysis, and SAS also provides the necessary tools.

References

Ades T (2009). Pair-wise meta-analysis and indirect comparisons: equally biased or equally meaningless? Slides presented at the ISCB Conference in Prague, available at http://www.iscb2009.info/RSystem/Soubory/Prez%20Monday/S01.2%20Ades.pdf

Bucher HC, Guyatt GH, et al. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* 50: 683–91.

Gleser LJ, Olkin I (1994). Stochastically dependent effect sizes. In: *The Handbook of Research Synthesis*. Eds Cooper H, Hedges LV. New York, Russell Sage Foundation: 339–355.

Janssen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Cappelleri JC (accessed Feb 2011). Interpreting indirect treatment comparisons & network meta-analysis for healthcare decision-making: Report of the ISPOR Task Force on Good Research Practices - Part 1. http://www.ispor.org/TaskForces/documents/Interpreting-Indirect-Treatment-Comparison-for-Decision-making-Part-1-FOR-COMMENT.pdf

Jones B, Roger J, Lane PW, Lawton A, Fletcher C, Cappelleri JC, Tate H, Moneuse P (2011). Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* **10**:523-531.

Lu G, Ades AE (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**:447–459.

Song F, Harvey I, Lilford R (2008). Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *Journal of Clinical Epidemiology* **61**:455–463.

Song F, Loke Y, Walsh T, Glenny AM, Eastwood AJ (2009). Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *British Medical Journal* **338**:b1147 doi:10.1136/bmj.b1147.