28

SYSTEMATIC REVIEWS AND META-ANALYSIS

OBJECTIVES

After reading this chapter, you should be able to:

1. Carry out a systematic review.

2. Complete the data-extraction process to provide data suitable for meta-analysis.

3. Calculate summary estimates of effect, evaluate the level of heterogeneity among study results and choose between using fixed- and random-effects models in your analysis.

4. Present the results of your meta-analyses graphically.

5. Evaluate potential causes of heterogeneity in effect estimates across studies.

6. Evaluate the potential impact of publication bias on your study results.

7. Determine if your results have been influenced by an individual study.

8. Deal with a variety of situations related to the types of data presented in, (or missing from) relevant studies.

9. Understand the important issues when carrying out a meta-analysis of observational studies.

28.1 Introduction

When making decisions about animal health interventions, we would like to use all of the information available in order to make the most informed decision. Unfortunately, the information in the literature is often inconclusive and conflicting. For example, the introduction of the use of recombinant bovine somatotropin (rBST) in the United States in 1994 initiated a substantial discussion of the potential effects of the use of the drug on the risk of clinical mastitis in dairy cows. If, in 1998, you carried out a review of all randomised clinical trials of rBST which reported risk ratios (or the data required to calculate a risk ratio) for clinical mastitis, you would have found 20 studies (representing 29 groups of cows) (Dohoo *et al*, 2003a). The point estimates of the risk ratio (*RR*) in those studies ranged from 0.67 (*ie* a reduction in risk) to 4.87 (a substantial increase in risk) (see Example 28.1). However, the effect was not statistically significant in 28 of the 29 groups studied. This might have led you to conclude that there was no effect of rBST on the risk of mastitis. Nonetheless, you might be left wondering if the variation in results was more than would be expected due to chance variation and what the power of each study to detect an effect was.

Similarly, if you carried out an evaluation of the effects of rBST on milk production (measured as 3.5% fat-corrected milk), you would have found data on 28 groups of cows in 19 different studies (Dohoo *et al*, 2003b). The point estimates ranged from a loss of 0.7 kg/day to a gain of 10.6 kg/day. Although there was a wide range of point estimates, the vast majority were over 3 kg/day and 23 of the 28 groups had statistically significant increases in production. Consequently, while it was clear that there was an effect, you might be interested in what the average effect was and why it varied from study to study.

If you wanted to carry out a more formal review of the available data on the effect of rBST on mastitis risk, there are 2 fundamental approaches which you could take: a narrative review or a systematic review (which might include a meta-analysis).

28.2 Narrative reviews

In situations in which there are very few studies (or the reviewer has chosen to only review a few studies), a review may take the form of a study-by-study report. In this case, each study is considered individually in order to subjectively take into account the unique circumstances of each study, and little effort is made to present an overall summary assessment of effect. If this was done thoroughly for the rBST data, it would identify the fact that each of the individual studies had very limited power to detect a moderate effect of rBST and the precision of each estimate of the RR was low. It is also likely that, with so much data available, you would like some form of summary estimate of the effect derived from all of the studies and a study-by-study report would not provide this.

A second approach would be to carry out a traditional narrative review in which studies are qualitatively assessed and the results subjectively combined into an overall conclusion. Some characteristics of traditional narrative reviews that make them less desirable than systematic reviews follow (Sargeant *et al*, 2006).

- They tend to be carried out by subject experts who may bring preconceived opinions to the process resulting in a biased review.
- They often do not have a structured methodology for identifying and assessing the relevant studies, leading to the possibility of selective inclusion of studies in support of

Example 28.1 Individual point estimates of risk ratio for effect of rBST on clinical mastitis
data = bst_mast

Twenty studies, containing data from 29 separate groups of cows had sufficient data to be able to calculate the risk ratio of the effect of rBST on clinical mastitis. The individual point estimates from each of the 29 groups were:

study	group	RR	study	group	RR	study	group	RR
1	1	0.83	6	11	1.00	15	21	1.19
1	2	0.91	7	12	0.96	15	22	1.26
2	3	1.08	8	13	0.95	16	23	1.40
3	4	1.30	8	14	1.31	16	24	0.67
3	5	0.90	9	15	1.45	16	25	1.11
4	6	1.75	10	16	1.02	17	26	4.87
4	7	1.45	11	17	1.40	18	27	2.60
4	8	0.83	12	18	1.80	19	28	4.00
4	9	1.35	13	19	1.73	20	29	1.37
5	10	2.50	14	20	1.91			

the reviewer's opinion.

- Small but well-designed studies may be omitted if they lack statistical power to produce statistically significant results.
- Inclusion criteria for studies are often not described in adequate detail for the reader to assess the thoroughness of the literature search.
- In deriving an overall estimate of effect, there is also a tendency to weight all studies equally, and as will be seen later, they should not all receive equal weight

As a consequence of the above limitations, narrative reviews should only be used to provide an overview of literature on a specific topic, not to guide treatment or prophylactic decisions.

28.3 Systematic Reviews

A recently published review outlines the steps involved in carrying out a systematic review (Sargeant *et al*, 2006). These are:

1. Specify the question to be answered
2. Lay out the review protocol
3. Find all of the studies
4. Determine which studies are relevant (requires inclusion and exclusion criteria)
5. Evaluate study quality
6. Extract the relevant data from each study
7. Summarise and synthesise the results (may include performing a meta-analysis)

Each of these is discussed below.

28.3.1 Specify the question

When specifying the question to be answered, you need to keep in mind what is most important from a clinical or animal-health policy objective, rather than letting data availability drive the study objective. It is often more desirable to address a more general question, which will broaden the eligibility characteristics for studies to be included in the review, rather than to address a very specific, but restrictive, question. For example, a review of the ability of *β*-blockers to reduce the short-term risk of myocardial infarction was based on studies in which 12 different drugs had been used (Freemantle *et al*, 1999) rather than focusing on a single specific drug. This enhanced the generalisability of the results.

Specifying the question, in addition to clarifying the intervention(s) to be considered will also involve specifying the specific outcome(s) considered, the comparisons to be evaluated (*eg* new treatment vs standard treatment or vs no treatment) and study designs to be included in the review.

28.3.2 Lay out the protocol

A systematic review should be both as objective as possible and sufficiently transparent so that a reader of the review should be able to duplicate it if they desired. This requires that a written protocol for the review be developed. This protocol corresponds to the 'Materials and Methods' section of a primary study and covers all of the steps (described below) in carrying out the review. Having a clear protocol will minimise the number of subjective decisions made during the review process.

28.3.3 Find the studies

The literature review on which a meta-analysis is based must be both complete and well-documented. The most commonly used approach to ensuring that all published studies are found is to carry out computer-based literature searches of the major electronic databases (*eg* Medline, Agricola, Index Veterinarius and the Veterinary Bulletin) and to follow this with a review of the reference lists in all of the papers identified through the computer-based search. Robinson and Dickersin (2002) give an example of an effective search strategy for searching for randomised controlled trials in Medline. The search process, including the names and date ranges of all databases searched along with the search strategy (*eg* keywords used) must be documented.

Finding articles published in conference proceedings and other forms (*eg* theses, non-indexed portions of journals) is more difficult than finding articles in the peer-reviewed published literature but necessary (Hopewell *et al*, 2002). While some databases of these resources are available, it is often necessary to identify conferences which were likely to have relevant publications in their proceedings and carry out a manual search of all proceedings for the time frame of interest.

One of the difficult issues to address is whether or not the review should include data from unpublished studies. The potential effects of publication bias are discussed in Section 28.8, but identifying and obtaining results from unpublished studies is a difficult task. In some cases, databases of funded research projects could be used to identify studies that have been conducted, but not published. Alternatively, personal contact with investigators working in the field might identify unpublished studies.

28.3.4 Determine if studies are relevant

The process of deciding whether or not studies are relevant involves specifying the inclusion and exclusion criteria for the review. Inclusion criteria include: the intervention(s) of interest, the populations(s) in which the studies can be carried out, the outcome(s) of interest, and the types of study to be included (many systematic reviews only consider randomised controlled trials, but this is not always applicable). Exclusion criteria may include factors such as publication in a language not accessible to the review team, publication prior to a specified date *etc*. The relevance of studies can usually be determined from the title and abstract and should be assessed independently by 2 or more members of the review team.

28.3.5 Evaluate study quality

The internal and external validity of each relevant study needs to be evaluated (with emphasis on internal validity). While inclusion criteria (described above) will play a role in ensuring the validity of studies considered (*eg* a meta-analysis might be based only on randomised controlled trials), other issues of study design (*eg* blinding, formal method of randomising treatment allocation, clear criteria for eligibility of subjects in the trial) must also be evaluated. A variety of scales and checklists have been developed (Jüni *et al*, 2001) and the list used will depend on the type of studies being evaluated (*eg* observational studies have different criteria than randomised controlled trials). The Cochrane Collaboration has developed a tool for assessing the risk of bias in studies which covers 6 domains: sequence generation, allocation concealment, blinding, incomplete data, selective reporting and other sources of bias. (See Cochrane Handbook for Systematic Reviews of Interventions for details (Higgins & Green, 2008).

Results of this quality assessment may be used in 1 of 3 ways. First, if a study does not meet all (or a subset) of the quality criteria, you might decide to exclude it from the meta-analysis. However, if very stringent criteria are set, you might end up excluding most studies.

The second approach is to evaluate study design issues and assign a quality score to the study. This quality score can be used to eliminate studies from consideration or to weight the studies in the meta-analysis (*ie* poor quality studies receive less weight when estimating the summary effect). The use of a quality scale introduces a degree of subjectivity to the meta-analysis and is not generally recommended (Greenland, 1994; Herbison *et al*, 2006; Higgins & Green, 2008).

The third approach is to record the key elements of the quality assessment and evaluate them as a source of heterogeneity (see Section 28.7) between studies. The quality assessment can also be used in a form of sensitivity analysis in which overall results are compared with those obtained from studies with defined subsets of quality characteristics.

28.3.6 Extract the relevant data

The layout and presentation of results in epidemiologic studies is highly variable. This is particularly true for observational studies, but it is even an issue when reviewing randomised controlled trials. The 2 fundamental pieces of information that you need from each study are the point estimate of the outcome(s) of interest and a measure of the precision of that estimate (SE or CI). In some cases, these are not presented directly, but sufficient data are available to allow you to compute the required information. For example, in the rBST studies referred to above, the primary outcome for most studies was a measure of productivity, but the number of cows in each study group which had one or more clinical cases of mastitis was also reported. From these data, the risk ratio for mastitis and its CI could be computed and used in the meta-analysis.

For outcomes measured on a binary scale (*eg* occurrence of clinical mastitis), you need to decide if you will extract and record a relative measure of effect (*eg* risk ratio—*RR*) or an absolute measure (*eg* risk difference—*RD*). It is generally more meaningful to use relative measures for summarising effects. The summary estimate can then be applied to specific populations in which the overall risk of disease is known (or can be estimated) to compute an absolute effect of the intervention. Regardless of which measure of effect is used, you should record the frequency of disease (*eg* risk) in the control group as this might be a source of heterogeneity of study results (see Section 28.7.4). In addition to data on the outcome of interest, bibliographic information and information on study characteristics (*eg* population, specifics of the intervention, length of follow-up *etc*) should be recorded.

Before starting the data-extraction process, you need to develop a template on which to record all of the fundamental information about the study, including any information required in the evaluation of the quality of the study or to evaluate as a possible cause of heterogeneity among study results. Given that data extraction is a complex process, it is desirable to carry out duplicate data extraction (*ie* data extracted independently by 2 investigators) followed by a comparison of the 2 datasets to identify and resolve any differences (Buscemi *et al*, 2006). When carrying out the data extraction, it is also important to watch for duplicate reporting of results. In some cases, data from an individual study might be published in multiple locations (*eg* a company report and a peer-reviewed journal publication) but must only be included in the meta-analysis once. Example 28.2 describes the literature review and data-extraction process for the meta-analysis of rBST. These data are used for all subsequent examples in this chapter.

Example 28.2 Literature review and data extraction for meta-analysis

The meta-analysis of the effects of rBST on dairy cattle productivity and health was carried out by an expert panel of the Canadian Veterinary Medical Association at the request of Health Canada. The data for the meta-analyses were obtained through the following process. A literature review of 3 electronic databases covering the period 1984 to 1998 identified a total of 1,777 references related to rBST. A review of the titles identified 242 manuscripts that potentially contained results from randomised clinical trials or were relevant reviews of the subject. These were all reviewed by the panel members and 60 identified as useful for the review. These were combined with 26 unpublished study reports provided as part of the company's submission in support of the request for registration of the drug. From all of these reports (n=86), 53 studies (representing 94 distinct groups of cows) were found to contain original data from randomised controlled trials. Estimates of effect on the various outcomes of interest were obtained and used in the meta-analyses.

Only data relating to milk production (3.5% fat-corrected milk) and the risk of clinical mastitis are presented in this chapter. A more detailed description of the methods used and estimates of effects on other parameters have been published (Dohoo *et al.*, 2003a; Dohoo *et al.*, 2003b)

28.3.7 Summarise and synthesise the results

Extracted data can be summarised and synthesised using qualitative or quantitative methods. A qualitative summary may involve tabular and/or graphical display of the key outcomes along with a narrative description of the studies. In situations in which there are few studies and/or the results from the studies are highly variable, a qualitative summary may be all that is warranted. In other situations, it is often desirable to compute an overall estimate of the outcome of interest and to quantitatively investigate why estimates of the outcome vary across studies. This quantitative assessment is called a meta-analysis and is the subject of the rest of this chapter.

28.4 META-ANALYSIS—INTRODUCTION

A meta-analysis has been defined as: "The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (cited in Dickersin & Berlin, 1992). It is a formal process for combining results from a number of studies that is being used increasingly in human medicine and, to a more limited extent, in veterinary medicine. Meta-analyses have been used most commonly to combine results from a series of controlled trials and this chapter will focus on that application. However, they can also be used to combine results from a series of observational studies (see Section 28.11) as was done in a meta-analysis of the effects of disease on reproductive performance in dairy cows (Fourichon *et al*, 2000). A more complete description of meta-analyses can be found in texts such as Egger *et al* (2001) or the online text published by the Cochrane Collaboration (Higgins & Green, 2008). The Cochrane Collaboration is an international organisation set up to help healthcare professionals make informed decisions through the use of systematic reviews of health research. A review of recent advances in meta-analysis methodologies has been published recently (Sutton & Higgins, 2008).

The objectives of a meta-analysis are to provide an overall estimate of an association or effect based on data from a number of scientific studies and to explore reasons for variation in the observed effect across studies. It accomplishes this by imposing a systematic methodology on the review process. Because it combines data from multiple studies, there is a gain in statistical power for detecting effects. When computing an overall estimate of effect, it takes into account both the individual study estimates and the precision of those estimates (standard errors) so that the results from each study are weighted appropriately.

Meta-analyses can be used to review existing evidence prior to making clinical or animal-health policy decisions, or as a precursor to further research by better quantifying what is already known, and identifying gaps in the scientific literature. A meta-analysis might be combined with a traditional narrative review and hence, should be thought of as complementary to that review process.

28.4.1 Meta-analysis—types of data

There are 3 types of data which can be used in a meta-analysis: summary estimate data, group data and individual patient (subject) data (IPD) (Table 28.1). Summary data are the most commonly used and consist of a point estimate of the effect of interest and some measure of its precision. For example, a set of studies might all report a risk ratio and its confidence interval for an effect of a treatment. If only summary data are available, the meta-analysis is restricted to using the measure(s) of effect reported. Group data consist of outcome data for each of the intervention groups (*eg* risk of 'cure' in the treatment group and the control group separately). For studies with binary interventions (treatment) and outcomes, it is usually possible to reconstruct the 2X2 table from which a variety of measures of effect can be computed. IPD are least frequently available and consist of the original data for each individual in the study. Summary estimate data can only evaluate the effects of study level variables (*eg* was the study blinded or not) as sources of heterogeneity. Group data can also include group-level covariates , although these are not usually important if study subjects have been randomly assigned to groups. IPD allow for the evaluation of study-, group- and individual-level variables (*eg*. age of study subjects) as a source of heterogeneity (see Section 28.7).

Table 28.1 Types of data used in meta-analyses

Data Type	Binary outcome	Continuous outcome
Summary estimate	Point estimate: RR, OR, RD, IR Estimate of precision: SE or CI	Point estimate: mean difference (MD) Estimate of precision: SE or CI
Group	Cell values for treated and control groups (from which various effects measures and their precision can be calculated)	Number, mean and standard deviation in each group from which the MD and its SE can be computed.
Individual patient data	Raw data—outcome value (0 or 1) and individual characteristics for each study subject	Raw data—outcome value (continuous) and individual characteristics for each study subject

Formulae for common measures of effect (*eg RR*, *OR etc*) and their standard errors and confidence intervals are presented in Chapter 6. If IPD are available, all of the data can be pooled into a single dataset and reanalysed, taking into account the clustered nature of the observations (within study) using methods outlined in Chapters 20-23. This is the most flexible approach to the analysis but these data are rarely available so it will not be considered further in this chapter.

28.4.2 Meta-analysis—process

There are multiple steps involved in carrying out a meta-analysis of results from a systematic review. These include:

- deciding whether to base the analysis on a fixed- or random-effects model
- computing a summary estimate of effect (if appropriate)
- presenting the data (usually graphically)
- evaluating possible reasons for heterogeneity of study results (*ie* why different studies produce different estimates)
- searching for evidence of publication bias and evaluating the influence that individual studies have on the outcome.

Each of these will be discussed in subsequent sections

28.5 FIXED- AND RANDOM-EFFECTS MODELS

A fundamental decision to be made in any meta-analysis is whether to use a fixed- or random-effects model. A fixed-effects model is based on the assumption that the effect of the factor being investigated is constant across studies and that any variation among studies is due only to random variation. On the other hand, a random-effects model assumes that the true study effect does vary across studies and the observed study effects reflect both this variation and random variation. Graphically, the studies can be represented as shown in Fig. 28.1.

28.5.1 Fixed-effects model

A fixed-effects model can be written as:

$$T_i = \theta + \varepsilon_i$$

Eq 28.1

where T_i is the effect measure (*eg* ln*RR*) from study *i*. (**Note** T_i is used to designate the study outcome instead of Y_i, which is generally used throughout this book for outcome variables, in order to distinguish between the measure of effect (T_i) and the outcome measured on individual study subjects Y_i).

θ is the overall treatment effect and ε_i is the error term for study *i* which is distributed as:

$$\varepsilon_i \sim N(0, V_i)$$

Eq 28.2

where V_i is the within-study variance for study *i* (V_i=[SE(T_i)]^2). This is assumed to be known and uncertainty about V_i is not part of the modelling process. Combining Eq 28.1 and Eq 28.2 shows that the distribution of the T_i is:

$$T_i \sim N(\theta, V_i)$$

Eq 28.3

Computing a summary estimate of effect

In order to compute a summary estimate of the overall effect, the individual study results must be weighted based on the precision of the estimates. The most commonly used procedure is **inverse variance weighting** where weights are computed as W_i=1/V_i This procedure is applicable for pooling results from models of continuous (linear regression, ANOVA) and discrete (logistic, Poisson regression) data. However, inverse variance methods might not work well when study sizes are small.

For binary data, alternative approaches based on the **Mantel-Haenszel** procedure or an approach attributable to **Peto** are available (Egger *et al*, 2001; Sweeting *et al*, 2004). The former is often preferable to the inverse variance approach when the data are sparse (*ie* where the outcome is a relatively rare event). The Peto method does not work well when treatment effects are large or when the intervention groups are seriously unbalanced (unequal sample sizes); however, the method has been extended for use with time-to-event data.

For continuous data there are 2 possible measures of effect size: the **mean difference** and the **standardised mean difference**. The mean difference is used when all studies measure the outcome on the same scale (which simplifies pooling the results) and the used weights are the inverse variance weights. Methods based on standardised mean differences are discussed in Section 28.10.2.

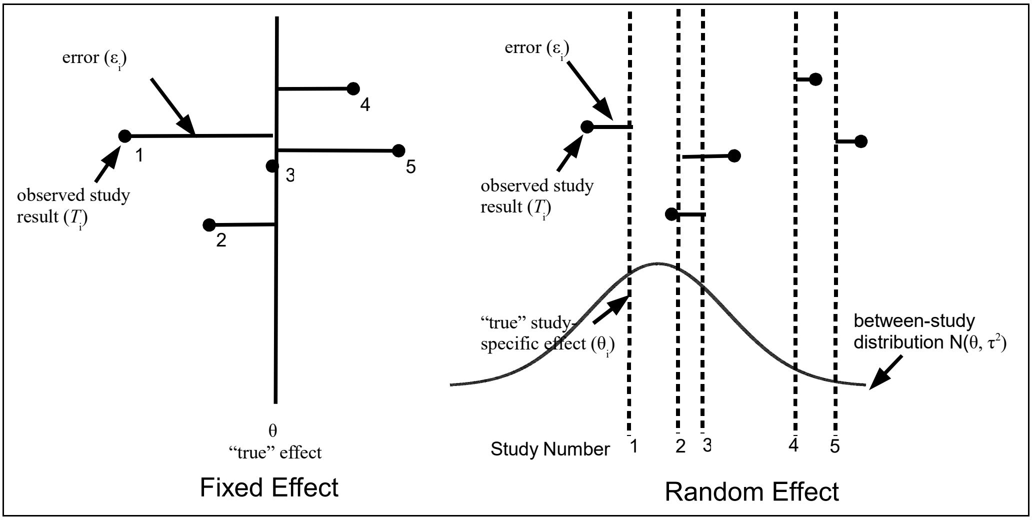

error (ε_i)
observed study result (T_i)
1
2
3
4
5
θ
"true" effect
Fixed Effect
error (ε_i)
observed study result (T_i)
"true" study-specific effect (θ_i)
between-study distribution N(θ, τ²)
Study Number 1 2 3 4 5
Random Effect

Fig. 28.1 Graphical representation of fixed and random-effects models

28.5.2 Random-effects model

A random-effects model assumes that a distribution (heterogeneity) of true treatment effects across studies exists, resulting in additional variability among study results. It is most common to assume that the study effects have a normal distribution so the random-effects model becomes:

$$T_i = \theta + u_i + \varepsilon_i \qquad \textit{Eq 28.4}$$

with: $u_i \sim N(0,\tau^2)$ and $\varepsilon_i \sim N(0,V_i)$ ***Eq 28.5***

where u_i is the random-effect for study i, and τ^2 is the between-study variance (heterogeneity). Combining Eq 28.4 and Eq 28.5 shows that the distribution of the T_i is:

$$T_i \sim N(\theta, V_i + \tau^2) \qquad \textit{Eq 28.6}$$

Random-effects models generally produce a point estimate of the summary effect that is similar to that obtained from fixed-effects models, but which has a wider confidence interval than a fixed-effects model (because the variance of the estimate is larger).

The simplest (and classical) analysis of a random-effects model estimates τ^2 by a method of moments (MM) and computes a summary estimate from the weights $W_i=1/(V_i+\tau^2)$ (DerSimonian & Kacker, 2007; DerSimonian & Laird, 1986). Recent alternative approaches, derived from statistical inference for mixed models (Chapter 21), include maximum likelihood (ML), restricted maximum likelihood (REML) and empirical Bayes (EB) methods. The MM, ML and REML estimate may be biased unless τ^2 is small while the EB estimate has been found to generally be accurate (Sidik & Jonkman, 2007). If IPD are available, mixed models (as described in Chapters 20-23) can be used for meta-analyses. In this case, a random slopes model which allows for the estimate of the treatment effect to vary across studies is used (see Chapter 21 for a discussion of random slopes models).

The advantages of a fixed-effects model are that it does not require the estimation of τ^2 nor are there any distributional assumptions about u_i. However, the assumption of a constant treatment effect across all studies is often not tenable, and ignoring between-study variation may lead to Type I errors (for the statistical significance of θ) and confidence intervals too narrow for θ. Consequently, random-effects models are now more commonly used. Results from fitting both fixed- and random-effects models of the effect of rBST on milk production are shown in Example 28.3.

28.6 Presentation of results

One of the most important outputs from a meta-analysis is a graphic presentation of the results with the most commonly used format referred to as a **forest plot** which displays the point estimate and confidence interval of the effect observed in each study along with the summary estimate and its confidence interval. Fig. 28.2 shows a forest plot for the effects of rBST on the risk of clinical mastitis and the elements of the plot are described in Example 28.4.

In some cases, it might be desirable to order the individual studies according to some criteria such as year of completion (to see if there is a trend over time) or quality score (to see if study quality affects the observed effects).

Example 28.3 Fixed- vs random-effects models

data = bst_milk, bst_mast

Both fixed- and random-effects models were fit to both the milk production data and mastitis data from the meta-analysis of the effects of rBST on dairy cow productivity and health. In all models, the inverse variance approach (Section 28.5) was used to assign weights to the study results.

Milk production (28 studies)

Method	Pooled estimate (kg/day)	Z	P	95% CI Lower	95% CI Upper
Fixed	4.465	28.078	0.000	4.153	4.777
Random	4.434	14.911	0.000	3.851	5.016

The *Q* statistic (Section 28.7) for heterogeneity was 79.9 with 27 degrees of freedom (P=0.000) giving strong evidence of heterogeneity among study results. Potential reasons for this heterogeneity will be explored in Examples 28.5 and 28.6. As expected, the point estimates for the summary effect were quite similar, but the random-effects model produced wider confidence intervals.

Based on the random-effects model, the estimate of the between-study variance was 1.42(SD=√1.42=1.2) suggesting that 95% of the effects of rBST should lie between 4.4-2*1.2=2.0 kg/day and 4.4+2*1.2=6.8 kg/day. Higgins I^2 (Section 28.7.2) was 66.2%.

Mastitis (29 studies)

Method	Pooled estimate (RR)	Z	P	95% CI Lower	95% CI Upper
Fixed	1.271	4.016	0.000	1.131	1.429
Random	1.271	4.016	0.000	1.131	1.429

The *Q* statistic for heterogeneity was 16.4 with 28 degrees of freedom (P=0.096) giving no indication of heterogeneity among study results. **Note** Because *Q* <df, the estimated between-study variance is zero and the results from the fixed- and random-effects models are identical.

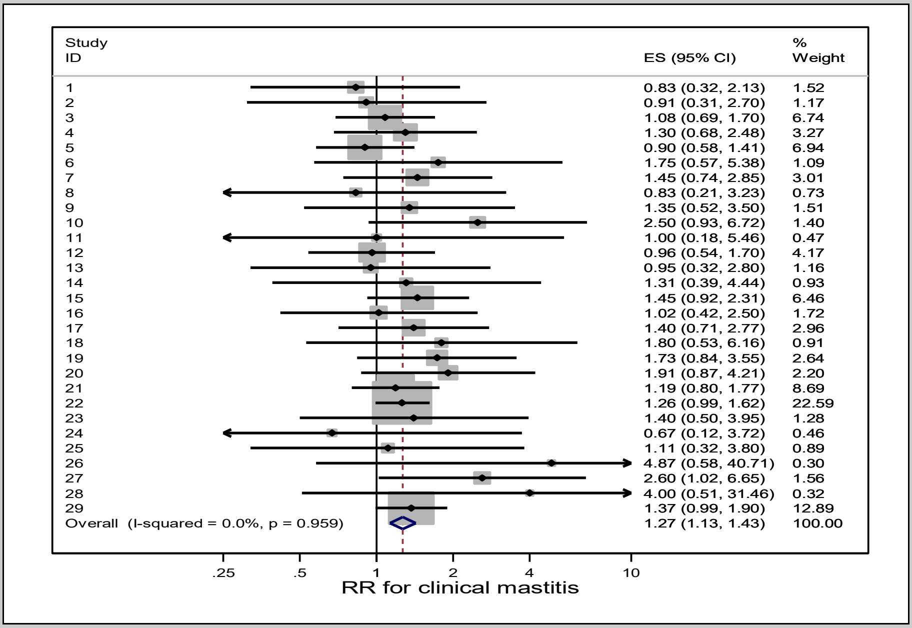
28.7 Heterogeneity

Heterogeneity refers to variability among study results (beyond random variation) and this should always be evaluated in a meta-analysis. Unfortunately, this is not always done—in a review of 34 meta-analyses carried out between 1999 and 2001, only 23 had any evaluation of heterogeneity (Petitti, 2001).

Example 28.4 Forest plot

data = bst_mast

Fig. 28.2 shows a forest plot of the risk ratios for the effect of rBST on the risk of clinical mastitis.



Study ID	ES (95% CI)	% Weight
1	0.83 (0.32, 2.13)	1.52
2	0.91 (0.31, 2.70)	1.17
3	1.08 (0.69, 1.70)	6.74
4	1.30 (0.68, 2.48)	3.27
5	0.90 (0.58, 1.41)	6.94
6	1.75 (0.57, 5.38)	1.09
7	1.45 (0.74, 2.85)	3.01
8	0.83 (0.21, 3.23)	0.73
9	1.35 (0.52, 3.50)	1.51
10	2.50 (0.93, 6.72)	1.40
11	1.00 (0.18, 5.46)	0.47
12	0.96 (0.54, 1.70)	4.17
13	0.95 (0.32, 2.80)	1.16
14	1.31 (0.39, 4.44)	0.93
15	1.45 (0.92, 2.31)	6.46
16	1.02 (0.42, 2.50)	1.72
17	1.40 (0.71, 2.77)	2.96
18	1.80 (0.53, 6.16)	0.91
19	1.73 (0.84, 3.55)	2.64
20	1.91 (0.87, 4.21)	2.20
21	1.19 (0.80, 1.77)	8.69
22	1.26 (0.99, 1.62)	22.59
23	1.40 (0.50, 3.95)	1.28
24	0.67 (0.12, 3.72)	0.46
25	1.11 (0.32, 3.80)	0.89
26	4.87 (0.58, 40.71)	0.30
27	2.60 (1.02, 6.65)	1.56
28	4.00 (0.51, 31.46)	0.32
29	1.37 (0.99, 1.90)	12.89
Overall (I-squared = 0.0%, p = 0.959)	1.27 (1.13, 1.43)	100.00

Fig. 28.2 Forest plot of studies of effects of rBST on risk of clinical mastitis in dairy cattle

In these plots, each horizontal line represents the results from a single study (or distinct group of cows within a study). Each line is labelled with a unique label (the group number). The length of the line represents the 95% confidence interval for the parameter estimate from the study. **Note** Some lines have been truncated at 10 or 0.25. The centre of the shaded box on each line marks the point estimate of the parameter from that study, and the area of the box is proportional to the weight assigned to the study in the meta-analysis. Studies with large boxes have had a strong influence on the overall estimate. The dashed vertical line marks the overall estimate of the effect. The <> at the bottom of the dashed line shows the confidence interval for the estimate of the overall effect. The solid vertical line marks the value where rBST would have no effect (*ie RR*=1).

As you can see, there was considerable variability among the individual study point estimates of the *RR* and only one of them was statistically significant (CI excludes 1). However, as seen in Example 28.3, this variability was not greater than what would be expected due to chance (given the generally small size of most of the studies). Study 22 had the largest influence on the summary result (*ie* greatest % weight).

28.7.1 Real vs artifactual heterogeneity

Heterogeneity may be real or artifactual. Real heterogeneity arises when there are true differences in treatment effects across studies. Artifactual heterogeneity arises when the difference is only due to study design issues, not to variation in the real treatment effect (Glasziou & Sanders, 2002). Study design issues, which might result in variation between observed effects across studies, include factors such as: duration of follow-up, reliability of the outcome measure (*ie* possibility for misclassification of the outcome), lack of blinding and/or compliance.

The choice of summary measure of treatment effect can also induce artifactual heterogeneity. For example, Table 28.2 shows some hypothetical data from 3 studies of a treatment. If the effect of the treatment is assessed using a risk ratio (*RR*), then all 3 studies show exactly the same treatment effect (2.0). However, if odds ratios (*OR*) or risk differences (*RD*) were used as a measure of effect, there would be substantial heterogeneity. In general, ratio measures of effect are considered more stable across studies than difference measures (Schmid *et al*, 1998).

Table 28.2 Hypothetical data from 3 studies showing that the choice of effect measure can influence whether or not there is heterogeneity among study results

Control[a]	Tx[a]	RR	OR	RD
0.1	0.2	2.0	2.3	0.1
0.2	0.4	2.0	2.7	0.2
0.4	0.8	2.0	6.0	0.4

[a] data are the proportion in each group with the outcome of interest

28.7.2 Clinical vs statistical heterogeneity

Another important distinction is between clinical and statistical heterogeneity. Clinical heterogeneity means that differences between populations studied (*eg* study selection criteria, disease severity, specifics of interventions) mean that ‘real’ differences are expected in the response (Egger *et al*, 2001). Statistical heterogeneity means that the variation between studies in the observed outcome (response to treatment) was more than would have been expected due to chance alone. If clinical heterogeneity is always expected, 2 important questions arise. First, is statistical assessment of heterogeneity warranted or should the focus be solely on quantification of the degree of heterogeneity? Second, is it reasonable to compute a single summary effect estimate given that the derived value is an average effect and may not apply to any specific population? Certainly, any summary effect measure must be interpreted with caution.

Despite the questions raised above, it is common to assess the statistical significance of heterogeneity and the most commonly used method is **Cochran’s *Q* statistic**. The formula is:

$$Q=\sum_{1}^{k} w_i(T_i-\theta)^2 \qquad \textbf{\textit{Eq 28.7}}$$

where w_i are the applied weights and θ is the pooled estimate (assuming a fixed effect). The null hypothesis tested is ‘no heterogeneity’ (*ie* τ^2=0 in Eq 28.5) and under this assumption it has a χ^2 distribution with *k*-1 df (where *k* is the number of studies). This test has relatively low power for detecting heterogeneity when the number of studies is small (Higgins & Thompson, 2002) or if the total number of study subjects is small or the study SEs vary considerably (Baujat *et al*, 2002). Consequently, the possibility of heterogeneity of effects should not be ruled out simply because the test yields a non-significant P-value. You might want to relax the P-value required for 'significant heterogeneity' (*eg* 0.1 instead of 0.05). Alternatively, you might want to evaluate the power of the test to detect heterogeneity among the studies you are evaluating (Hedges & Pigott, 2001; Jackson, 2006). If there is any evidence of heterogeneity, potential causes of that variability should be investigated (see Section 28.7.3).

The level of heterogeneity in a meta-analysis can be quantified using **Higgins I^2** which computes the proportion of variance between studies that is due to heterogeneity as opposed to chance (Higgins & Thompson, 2002; Higgins *et al*, 2003).

$$I^2=[Q-(k-1)]/Q * 100\% \qquad \textbf{\textit{Eq 28.8}}$$

The adjectives: low, medium and high were originally assigned to values of I^2 of 25%, 50% and 75%, respectively (Higgins *et al*, 2003), although an evaluation of possible causes of heterogeneity should be undertaken whenever the I^2 is greater than 25%.

28.7.3 Evaluation of heterogeneity

There are several possible approaches to evaluating heterogeneity and these include:

- subgroup analyses
- stratified analyses
- graphical assessment
- meta-regression.

Each of these will be discussed below.

Subgroup analysis
It may be possible to identify a specific subgroup of studies defined by a characteristic of interest and to focus attention on that subgroup. However, results from a specific subgroup should be interpreted with caution. As a hypothetical example, consider a meta-analysis of 11 studies of oral calcium supplementation at the time of calving for the prevention of milk fever. An overall beneficial effect (*RR*=0.6) was observed but significant heterogeneity was present and breed of cow appeared to be a contributing factor. If most studies (*n*=10) were carried out in Holsteins but the single study carried out in Jerseys found no significant beneficial effect (*RR*=1.0), what advice would you provide to a Jersey herd owner? Provided there is no biological basis for expecting a substantial difference in the treatment effect between Jerseys and Holsteins, the best estimation of effect in any particular subgroup is provided by considering all of the evidence rather than just looking at the data from that subgroup (this is referred to as **Stein's paradox** (Egger *et al*, 2001). Consequently, the advice to the Jersey herd owner should probably be that the therapy is effective. In general, results from specific subgroups should only be considered if the intent to evaluate that subgroup was clearly spelled out in the systematic review protocol (Section 28.3.2) (Higgins *et al*, 2002).

Stratified analysis
In a stratified analysis, the data are stratified according to a factor (which should be specified in the study protocol) thought to influence the treatment effect, and a separate meta-analysis carried out in each of the strata. The disadvantage to this approach is that individual strata might

contain relatively few studies. The statistical significance of the difference between 2 strata can be computed using a standard Z test statistic:

$$Z=\frac{\theta_1-\theta_2}{\sqrt{SE(\theta_1)^2+SE(\theta_2)^2}} \qquad \textit{Eq 28.9}$$

where θ_1 and θ_2 are the effects estimates in the 2 strata. (**Note** the problem of multiple comparisons must be avoided). Based on a fixed-effects assumption within each stratum, Cochran's Q statistic can be used to compute a test for homogeneity across strata (null hypothesis is that there are no strata effects).

$$Q_B=Q_T-\sum Q_s \qquad \textit{Eq 28.10}$$

where Q_T, and Q_S, are Cochran's Q statistics for the full data and stratum *s*, respectively. Q_B can be compared to a χ^2 distribution with *S*-1 *df* where *S* is the number of strata. However, this test statistic is only valid if there is no significant residual heterogeneity within any of the strata. Example 28.5 presents a stratified (by parity group) meta-analysis of the effects of rBST on milk production.

Graphical assessment of heterogeneity

Several types of plot can be used to evaluate the level of heterogeneity and the possible contribution of specific factors to the observed heterogeneity. A **Galbraith plot** plots the *Z* statistic (Z_i=T_i /SE(T_i)) from each study against the inverse of its SE (1/SE). The slope of the resulting line is the overall (fixed-effect) estimate, and lines ±2 units from this line should encompass 95% of observations if there is no significant heterogeneity. The plot can also be used to identify outlying points contributing substantially to the Q statistic. Fig 28.3 shows a Galbraith plot for the rBST—milk production data (8 of 28 (29%) of the observations lie outside the ± 2 unit lines).

Simple **scatter plots** of the effect size vs a factor suspected of contributing to heterogeneity can be useful for evaluating that relationship. If the effect measure is a ratio measure (*eg RR* or *OR*), then the log of the effect should be plotted. Fig. 28.4 shows a scatter plot of the log *RR* of clinical mastitis vs the daily dose of rBST with the size of the plotted points proportional to their weights and points labelled with study identifiers to help identify outlying observations. There is very slight evidence of a trend in effect (log *RR*) with increasing daily dose.

28.7.4 Meta-regression

The most flexible approach to evaluating causes of heterogeneity is **meta-regression.** A meta-regression is a weighted regression of the observed treatment effects against study-level predictors (with inverse variance weights used most commonly for the weightings). If the number of studies is limited, factors might be investigated one at a time, or if there are sufficient data, a multivariable regression model could be built.

As with meta-analyses, both fixed- and random-effects models are possible (Higgins & Thompson, 2004). A fixed-effects model assumes that the factors in the model completely explain the between-trial variance (*ie* the predictors completely explain the between-study variance). This is usually an unjustifiable assumption and often leads to Type I errors. Consequently, fixed-effects models should not be used.

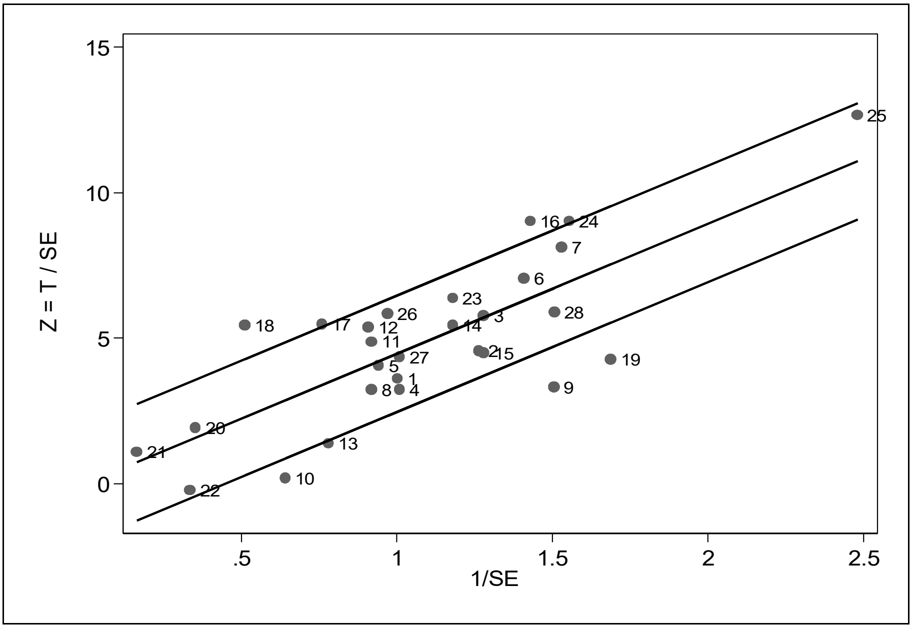
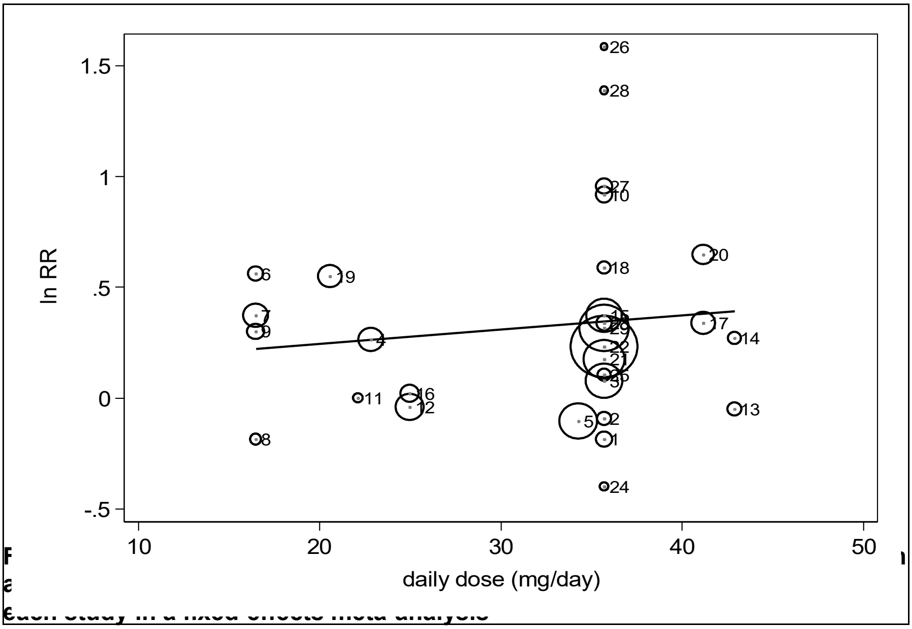


Fig. 28.3 Galbraith plot for assessing heterogeneity in the effect of rBST on milk production



F n
a
each study in a fixed effects meta analysis

Example 28.5 Stratified meta-analysis

data = bst_milk

Separate meta-analyses of the effect of rBST on milk production (kg milk/day) were carried out for each of the 3 parity groups: (primiparous, multiparous and no separation by parity) (*ie* studies which did not stratify on the basis of age).

Parity group	Number of studies	Estimate	Confidence Interval		Heterogeneity P	Higgin's I^2
No separation by parity	15	4.916	4.505	5.327	<0.001	69.6 %
Multiparous	7	4.361	3.700	5.022	0.68	0 %
Primiparous	6	3.300	2.608	3.992	<0.01	64.9 %
All data together	28	4.465	4.153	4.777	<0.001	66.2%

Parity seems to account for some of the heterogeneity among studies (cows in different age groups respond differently to rBST), but the results are not clear cut. Within the groups of multiparous cows, there was no longer any evidence of heterogeneity. However, there was still heterogeneity among the studies based solely on primiparous cows. You might have expected groups in which data from all parities were combined to have an effect intermediate to the other 2 groups, but this was not the case. However, the number of studies within each group was quite small, so the summary effects must be interpreted with caution. Because of the heterogeneity in some of the groups, the Q_B should not be used to compare strata.

A random-effects meta-regression model extends the random-effects model (Eq 28.4) by adding predictors; *eg* a model with a single predictor can be written as:

$$T_i=\theta+u_i+\beta_1 X_{1i}+\varepsilon_i \qquad \textbf{\textit{Eq 28.11}}$$

with:

$$u_i \sim N(0,\tau^2) \qquad \text{and} \qquad \varepsilon_i \sim N(0,V_i) \qquad \textbf{\textit{Eq 28.12}}$$

where u_i are the random-effects for each study and τ^2 is the between-study variance. Estimation of τ^2 can be be based on the same methods as for Eq 28.4 (*ie* method of moments, ML, REML and EB). ML estimates may be biased downward because the ML estimation procedure does not take into account the degrees of freedom used in estimating the fixed effects (Sidik & Jonkman, 2007). Example 28.6 shows a meta-regression of the effects of rBST on milk production on parity group, company and daily drug dosage.

There are a number of issues to be considered when carrying out meta-regression analyses (Thompson & Higgins, 2002). First, it must be recognised that meta-regression analyses are observational studies, even if the individual studies in the meta-analysis were randomised controlled trials. Consequently, the role of confounding and intervening factors needs to be considered. For example, daily dose was one factor examined in Example 28.6. If the breed of cow in the study influenced what dose of rBST they were given and also affected the milk production response to treatment, then breed will be a confounding variable that needs to be controlled.

Second, the significance of individual predictors needs to be carefully considered. Knapp and Hartung (2003) introduced a variance estimator that produced CI with better coverage than standard estimates. However, when there are few studies, even this approach is too conservative and a non-parametric permutation approach may be preferred (Harbord & Higgins, 2008).

There is also the issue of multiple comparisons. Meta-regressions may evaluate a large number of predictors, perhaps based on a fairly small sample of studies. This greatly increases the probability of finding one or more significant associations due to chance alone (Higgins & Thompson, 2004). Some adjustment (*eg* Bonferroni or an adjustment based on a permutation approach to computing P-values (Harbord & Higgins, 2008)) for the number of predictors being evaluated may be necessary.

Finally, the potential for ecological fallacies must be considered (see also Chapter 29). Predictors in a meta-regression are study-level values and these may represent study-level averages for individual study subject-level characteristics (*eg* average age of cows in the study). A relationship observed at the study level may not be true at the study subject level. Meta-analyses based on IPD are much better suited for evaluating the effects of individual-level characteristics.

28.7.5 Underlying risk as a cause of heterogeneity

If the outcome of interest is binary, one potential cause of heterogeneity that deserves special consideration is the **underlying risk**, as measured by the risk in the control group. In each study, the risk of disease in the control group is a reflection of the overall risk in the population being studied. It is important to address the question—is the treatment more or less effective when disease is rare vs common? This issue can be examined graphically using a **L'Abbé plot**

Example 28.6 Meta-regression for evaluating causes of heterogeneity
data = bst_milk

A meta-regression was carried out to evaluate the effects of parity group, company (manufacturer of the product) and daily dosage on the effects of rBST on milk production.

REML estimate of between-study variance

Number of obs = 28
tau2 = 1.376
Higgins I^2 = 64.29%
Model F(6,21) = 1.47
Prob > F = 0.2351

	Coef	SE	t	P>t	95% CI	
par - multip	-0.754	0.816	-0.920	0.366	-2.451	0.944
par - primip	-2.034	0.884	-2.300	0.032	-3.872	-0.195
company=2	-0.832	2.312	-0.360	0.723	-5.640	3.976
company=3	0.238	1.058	0.220	0.824	-1.963	2.438
company=4	1.874	1.347	1.390	0.179	-0.928	4.676
daily dose	0.028	0.061	0.450	0.654	-0.099	0.155
constant	3.922	1.657	2.370	0.028	0.476	7.367

The overall model was not significant (P=0.235). Parity group was borderline significant (P=0.094). Compared with studies with all age groups combined (the baseline category) studies in multiparous cows had similar effects but a smaller effect was seen in studies of first parity animals only (primiparous). The parity effect was similar to those seen in the stratified analysis (Example 28.5). The other 2 factors were not significant predictors. The τ^2 value (1.38) is the estimate of the between-study variance after adjustment for the predictors in the meta-regression.

which plots the risk in the treated group vs the risk in the control group (Song, 1999). If there is little heterogeneity, the points will cluster around a line corresponding to the pooled treatment effect. Fig. 28.5 shows a L'Abbé plot for the rBST—clinical mastitis data with the size of the points proportional to the sample size of the study and studies labelled to help identify outliers. The graph shows no evidence that the treatment effect varies with the underlying risk.

Underlying risk seems like an ideal candidate to consider in a meta-regression. Unfortunately, there is a structural dependence between the underlying risk (risk in the control group) and the risk ratio because the latter includes the former in its calculation. Studies with a low risk in the control group are more likely to have higher risks in the treatment group (or vice versa) due to random variation. As a result of this structural dependence, standard meta-regression methods are not a good option for evaluation of the effect of underlying risk. A recent study (Dohoo *et al*, 2007) compared 3 methods (both Bayesian and frequentist) for evaluating the effects of underlying risk. The overall conclusions were:

- if underlying risk does contribute to heterogeneity, the estimate of the intervention (treatment) effect from an ordinary random-effects meta-analysis will be biased , but the bias is not generally large,
- one of the 3 models was generally recommended (because it required fewer assumptions),
- Bayesian methods were very flexible and provided direct estimates of the SE of the predictors
- frequentist methods worked better if there were few studies.

The overall conclusion was that it was probably reasonable to start with a standard meta-regression and use one of these more complex methods as a final step if there was evidence that underlying risk explained any of the heterogeneity.

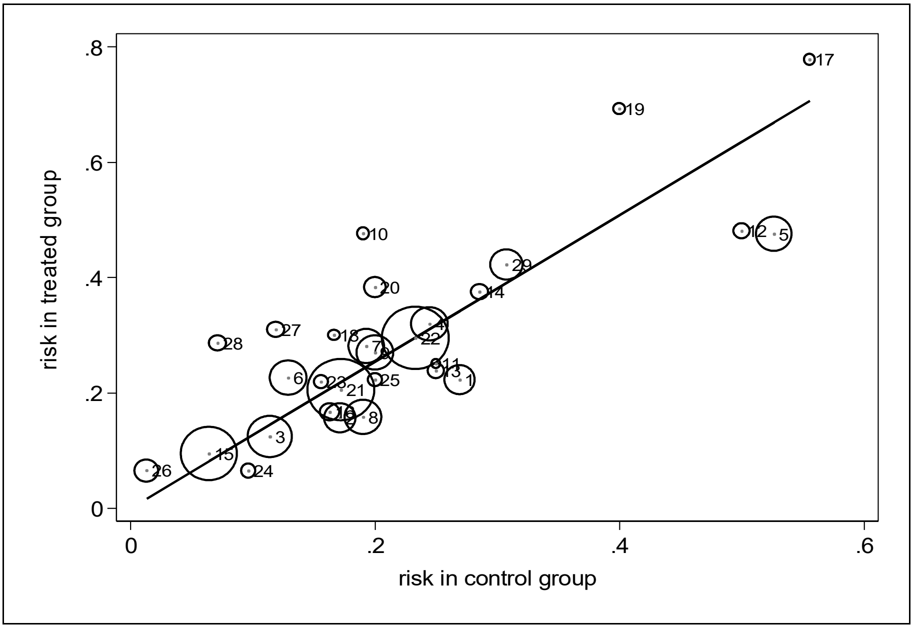


Fig. 28.5 L'Abbé plot for the rBST – clinical mastitis data

28.8 PUBLICATION BIAS

When carrying out a meta-analysis, you need to consider whether it is likely that there are studies that have been completed, but for which the results have not been published. Study results that are not statistically significant or which are unfavourable to the sponsor of the study might be less likely to be published than significant, favourable results (Dickersin, 1997). Consequently, published studies may represent a biased subset of the total body of work on the subject (Hopewell *et al*, 2007). Unfortunately, it is often very difficult to obtain unpublished study results. However, if you have any indication that unpublished results constitute a substantial portion of data available, then you should make an effort to obtain them. On the other hand, one argument against including unpublished results in a meta-analysis is that those results have not been peer reviewed and thus, do not have one of the key components in assuring data quality.

There are 3 general approaches to dealing with the problem of publication bias. The first, as described above is to contact investigators directly to obtain unpublished results, or to at least determine how many unpublished results there are. A second approach is to estimate how many studies with 'null' results (*ie* no observed effect) would have to exist before a summary effect from your meta-analysis would become non-significant. This approach is less recommended because it focuses on hypothesis testing (is there an effect or not?) rather than on estimating the magnitude of the overall effect.

The third approach is based on an evaluation of the relationship between study results and their precision. A **funnel plot** displays each study's estimated effect plotted against either its SE or its inverse (1/SE). If publication bias is a problem, there will likely be a number of studies with large effects and large SEs but an absence or shortage of studies with large standard errors and small or no effects. For example, Fig. 28.6 in Example 28.7 shows a funnel plot from a meta-analysis of the effects of anthelmintic treatment on milk production in dairy cows (Sanchez *et al*, 2004). There appears to be many more studies with large SE and positive treatment effects than comparable studies with negative or null treatment effects. This suggests that publication bias may be a problem.

There are a number of statistical tests based on the principle of the funnel plot. These evaluate the relationship between study results and their SEs using a rank correlation (Begg's test: Begg & Mazumdar, 1994) or a linear regression approach (Egger's test: Egger *et al*, 1997). Both tests standardise the observed effect sizes prior to evaluation of the association with the SE. Neither test is very sensitive if the number of studies is small (*eg* <20) (Sterne *et al*, 2000) although, in general, Egger's test is more powerful at detecting publication bias. (This does not appear to be the case in Example 28.7.) Both tests may also produce false positive results in situations in which there are large treatment effects, few events per trial or all trials are of similar size. If either test is significant, publication bias might be influencing your results. However, the tests are only appropriate when you expect either a positive or negative effect from the intervention (not both). If either positive or negative effects are of equal interest, any publication bias would produce a 'gap' in the 'middle' of the funnel (studies with null effects are less likely to be published) which would not likely be detected.

A 'trim and fill' method (Duval & Tweedie, 2000) of assessing the effect of possible publication bias is based on the following steps.

- 'Trim'—Produce a funnel plot and then sequentially omit studies until it is considered symmetrical

Example 28.7 Funnel plot for evaluation of publication bias

data = meta_parasite

A meta-analysis of the effects of anthelmintic treatment on milk production in dairy cows was carried out using data from 79 study groups from 55 studies. This evaluation of publication bias was limited to studies (*n*=18) which measured the treatment effect on 305 day actual milk production (excluding 2 very small studies with very large SE).

Fig. 28.6 shows a funnel plot in which it appears that there is a shortage of studies with a null or negative treatment effect and a large SE.

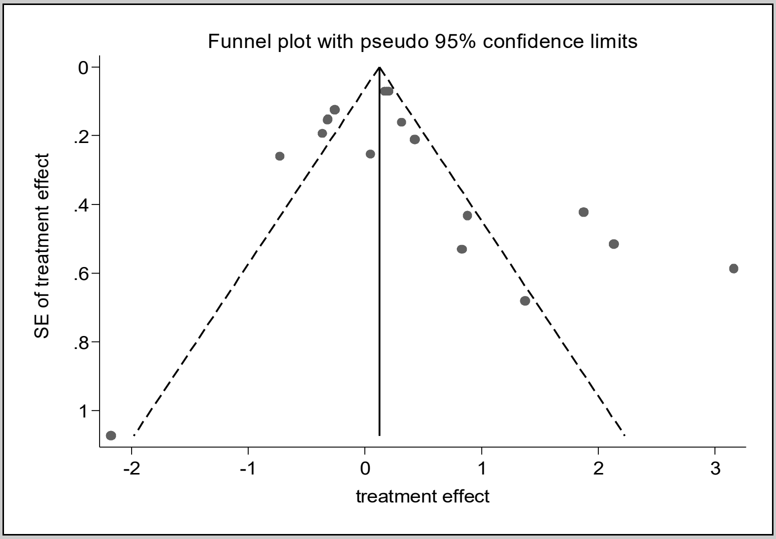


Fig. 28.6 Funnel plot

Begg's test of publication bias has a P-value of 0.068 while Egger's has a P-value of 0.342. The large discrepancy in the significance of the 2 tests is probably due to the small sample size included in this evaluation.

- Determine the centre of this new 'symmetrical' plot (*ie* a new estimate of the treatment effect)
- 'Fill'—Replace the omitted studies along with their 'counterparts' (hypothetical studies of the same SE but on the other side of the centre line
- Redo the meta-analysis using the original data plus the new hypothetical studies

This provides an estimate of what the treatment effect would be if all studies had been published. Example 28.8 shows a "trimmed and filled" funnel plot and shows how the estimated treatment effect is substantially reduced by the addition of the hypothetical studies.

Methods of evaluating publication bias are an active area of research and some recent areas of investigation are described in (Sutton & Higgins, 2008). However, we caution that the assessment of funnel plots is subjective and asymmetry is not easy to detect (Terrin *et al*, 2005). In addition, factors other than publication bias can produce asymmetry, so overly strong interpretation of funnel plots and tests of publication bias should be avoided.

Example 28.8 'Trim and fill' evaluation of publication bias
data = meta_parasite

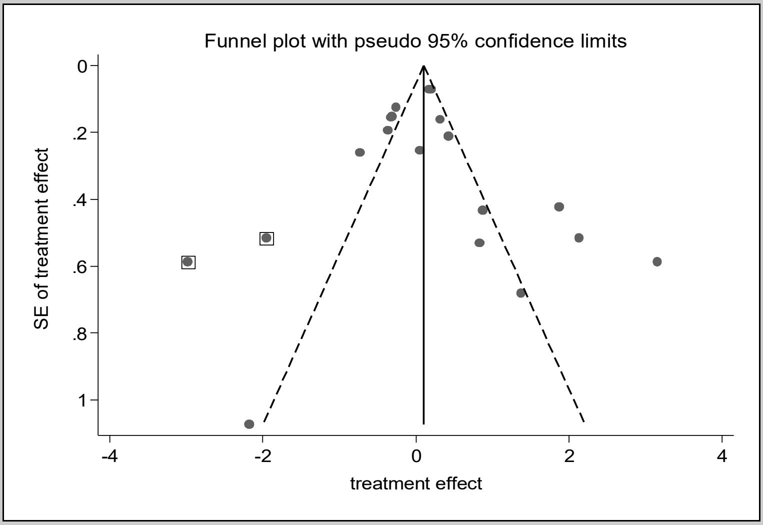


Fig. 28.7 Trim and fill evaluation of publication bias

Duval and Tweedie's 'trim and fill' method suggests that 2 additional studies (identified as points with boxes around them) need to be added to bring the funnel plot back to symmetry (Fig. 28.7). However, if these 2 studies are added the estimated treatment effect (random-effects model) drops from 0.252 (P=0.024) to 0.123 (P=0.316)

28.9 Influential studies

As in most regression-based models, it is important to determine if individual studies are having a profound influence on the summary estimate derived from a meta-analysis. If they are, you need to determine whether or not this is warranted. It might well be that one study was much larger than the others and consequently provides a much more precise estimate of the effect. In this situation, you need to evaluate that study to determine if it was of sufficiently high quality that you can accept the results.

One way to evaluate the effects of individual studies is to sequentially delete the studies from the meta-analysis and determine how the estimate of the summary effect changes (Example 28.9). The revised point estimates can all be plotted in an influence plot (see Fig. 28.8).

28.10 Outcome scales and data issues

Published manuscripts vary substantially in how much data they provide and how they present them. This gives rise to a number of data related issues which include:

- methods for computing standard errors (SE),
- dealing with continuous outcomes that may be measured on different scales,
- combining data from studies that use continuous and dichotomous outcomes,
- imputing missing variance estimates,

Example 28.9 Influential studies
data = bst_mast

An influence plot was generated to determine the effect of removing individual studies from the meta-analysis of rBST on the risk of clinical mastitis.

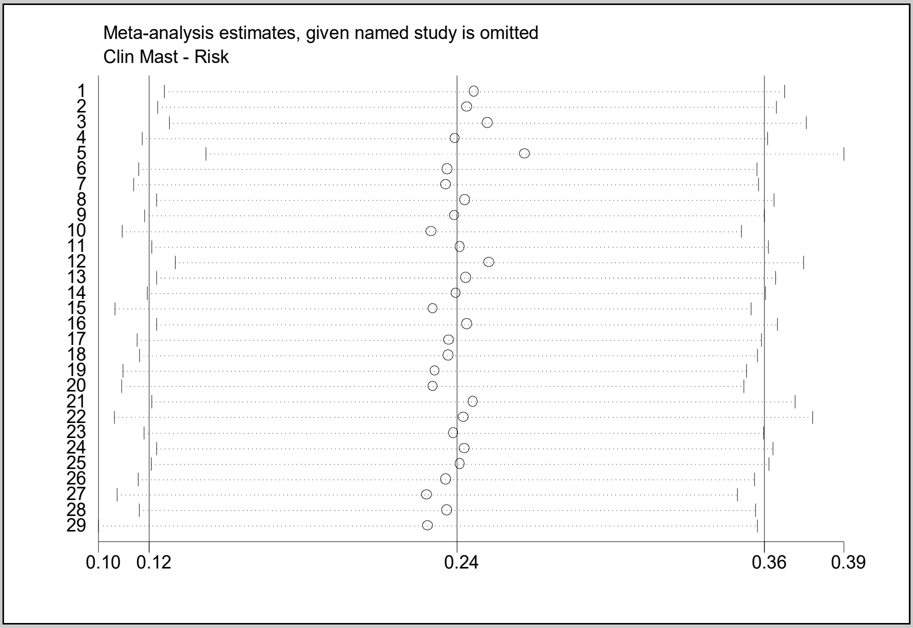


Fig. 28.8 Influence plot

No individual study (group of cows) had an undue influence on the summary effect estimate. Omitting study 5 had the largest effect and in this case the ln(*RR*) rose from 0.24 to 0.27 (equivalent to a rise in *RR* from 1.27 to 1.31). This is a relatively small change, indicating that no individual study had a particularly large influence on the summary *RR* estimate.

- imputing 2X2 table cell frequencies, and
- dealing with sparse data.

All of these will be dealt with below. However, regardless of the method(s) actually used, it is of paramount importance to ensure that any data modifications undertaken are biologically sensible. For example, if data from different scales are combined, it is necessary that they measure the same effect. Similarly, if data values are imputed, one should always check that the resulting values are reasonable and do not end up as influential values in the analysis.

28.10.1 Methods of computing standard errors

If a study reports a confidence interval but not a SE, the SE can be estimated as

$$SE=\frac{(\mathrm{UL}-\mathrm{LL})}{2\mathrm{Z}_{1-\alpha/2}} \qquad \textbf{\textit{Eq 28.13}}$$

where UL and LL are the upper and lower limits of the CI respectively. For a 95% CI,

$Z_{1-\alpha/2}=1.96$. (**Note** for small sample sizes a *t* statistic should be used instead of a *Z* statistic). For ratio measures (*eg* $\theta=RR$ or *OR*), the SE must be determined for $\ln(\theta)$ and it is estimated from the CI for $\ln(\theta)$.

Occasionally, a study reports grouped data or raw values, and it is necessary to compute the effect measure of interest and its SE. For dichotomous outcomes, formulae for measures of association (*RR*, *OR*, *RD*) and their SE are presented in Chapter 6. For continuous outcomes, the main measures which you may need to compute follows.

$$MD_i = m_{1i} - m_{2i}$$

Eq 28.14

where m_{1i} and m_{2i} are the mean values in groups 1 and 2, respectively, in study *i*.

$$\mathrm{SE}(MD_i) = \sqrt{\frac{SD_{1i}^2}{n_{1i}} + \frac{SD_{2i}^2}{n_{1i}}}$$

Eq 28.15

where the *SD*s and *n*s are the group specific standard deviations and sample sizes for study *i*. For some computations (see Section 28.10.2) it is necessary to compute a pooled *SD* (s_i).

$$s_i = \sqrt{\frac{(n_{1i}-1)*SD_{1i}^2 + (n_{2i}-1)*SD_{2i}^2}{n_{1i}+n_{2i}-2}}$$

Eq 28.16

28.10.2 Continuous outcomes measured on different scales

You sometimes encounter studies which have evaluated comparable outcomes, but which have used different scales to do so. For example, when carrying out a meta-analysis of studies into the effects of monensin in dairy cattle, Duffield *et al* (2008) needed to be able to include studies which measured metabolites on different scales from which there was no mathematical conversion possible. The solution to this problem is to compute standardised mean differences which, for each trial, expresses the treatment effect relative to the variability observed in the outcome in the trial. The resulting measure is often referred to as an **effect size** (ES) and there are 3 common methods of computing it.

Cohen's *d* relates the mean difference to the pooled *SD* of the 2 groups.

$$d_i = \frac{m_{1i}-m_{2i}}{s_i} \qquad \mathrm{SE}(d_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{d_i^2}{2(n_{1i}+n_{2i}-2)}}$$

Eq 28.17

where s_i is the pooled *SD* (Eq 28.16).

Hedges' adjusted *g* is a similar measure but includes a small sample adjustment.

$$g_i = \frac{m_{1i}-m_{2i}}{s_i}\left(1 - \frac{3}{(4(n_{1i}+n_{2i})-9)}\right) \qquad \mathrm{SE}(g_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{g_i^2}{2(n_{1i}+n_{2i}-3.94)}}$$

Eq 28.18

Glass's Δ scales the difference by the *SD* in the control group. It is the preferred measure if the intervention affects both the mean value of the outcome and its variability.

$$\Delta_i = \frac{m_{1i}-m_{2i}}{SD_{2i}} \qquad SE(\Delta_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{\Delta_i^2}{2(n_{2i}-1)}}$$

Eq 28.19

Methods of combining results when studies report outcomes using a mixture of raw and log-transformed scales have recently been published (Higgins *et al*, 2008).

28.10.3 Combining continuous and dichotomous outcomes

In some situations, it may be necessary to combine outcomes measured on a continuous scale in some studies and a dichotomous scale in others. One approach to this problem is to compute an effect size from the 2x2 tables in the studies with a dichotomous outcome (*ie* convert the *RD*, *RR* or *OR* into an ES). Seven approaches to this computation have been evaluated (Sánchez-Meca *et al*, 2003) with the most widely applicable being one attributed to Cox in which a ln(*OR*) is converted to an effect size ($d_{\mathrm{Cox}}=\ln(OR)/1.65$) with its associated SE=SE(ln(*OR*))/1.65.

An alternative approach is to dichotomise the results from the studies based on a continuous outcome based on a selected cutpoint. The disadvantages of this approach are that it only works if all studies with continuous outcomes used the same scale (or you need to select a cutpoint for each scale used) and any dichotomisation involves a loss of information.

28.10.4 Imputing missing variance estimates

If studies with continuous outcomes fail to report either the SE of the mean difference, or the *SD*s for the study groups, you need to come up with an estimate of the precision of the mean difference if the study is to be included in the meta-analysis. There are multiple approaches to dealing with this problem (Thiessen Philbrook *et al*, 2007; Wiebe *et al*, 2006). If either a P-value or a test statistic for the difference between the 2 groups is given, the SE can be computed from standard statistical formulae. If only a range was reported for the P-value (*eg* P<0.05), you can take a conservative approach and use the largest possible P-value.

An alternative approach is to 'borrow' a SE estimate from other studies. This can be done by choosing the largest SE reported (conservative), using the mean SE from all other studies or imputing the SE based on study characteristics (usually done using a linear regression model). Using a 'borrowed' SE has been found to perform acceptably (Furukawa *et al*, 2006) and is preferable to omitting studies with missing SEs from the meta-analysis.

28.10.5 Imputing 2x2 table cell frequencies

Some meta-analysis procedures for dichotomous outcomes require 2x2 table cell frequencies, not just an effect measure (*eg* ln(*OR*)) and its SE. For example, without cell frequencies, the only weighting method which can be used is the inverse variance approach and it has some limitations (see Section 28.5.1). Algebraic methods of imputing cell frequencies from an effect measure (*RD, RR* or *OR*), their CI (or variance) and their sample size have been reported (Di Pietrantonj, 2006). The accuracy of the estimation depends on the precision of the CI (or variance) reported and the width of the CI (wider CI leads to better imputation). It also depends on the number of significant digits to which the CI is reported, but is generally adequate if reported to 2 decimal places.

28.10.6 Dealing with sparse data

You may encounter situations in which there are zero events in one or both of the intervention groups (*eg* if adverse reactions are one of the outcomes assessed in a meta-analysis of RCTs). If there are zero events in both groups, the study can be ignored because it contains no information. If there is a zero in one group, the impact will depend on the method of pooling. If inverse variance weights are used, the ln*OR* or ln*RR* are undefined. Mantel-Haenszel weighting may not be possible but the Peto method is not affected. Sweeting *et al* (2004) review and evaluate different continuity adjustments that can be used to deal with the problem of sparse data (details beyond the scope of this text) and Rücker *et al* (2009) propose the use of arcsine differences as an alternative.

28.11 Meta-analysis of observational studies

While meta-analyses have more commonly been used for combining results from RCTs, meta-analyses of observational studies are becoming more common and are equally important (Dickersin, 2002; Egger *et al*, 2001). One notable example is the 1964 Surgeon General's Report on Smoking and Health (Surgeon General, 1964) in which the effect of smoking on cancer risks from 7 cohort studies were evaluated. There are a number of reasons why the use of meta-analysis lags behind that of RCTs (Dickersin, 2002).

- There has been less research into meta-analysis methods applicable to observational studies.
- There is not yet a register of observational studies being conducted (in veterinary medicine there is no register of RCTs either).
- Efforts to standardise methods of reporting of observational studies (*eg* the STROBE Statement (von Elm *et al*, 2007)) are very new.
- Causal criteria need to be considered (see Section 1.10). A meta-analysis may enhance the statistical evidence and an evaluation of heterogeneity gives insight into the consistency of results, but other criteria also need careful consideration.

Observational studies are prone to a wide range of biases and there is a risk that a meta-analysis may lend spurious precision to questionable results (Egger *et al*, 1998). Consequently, the focus of a meta-analysis of observational studies should be an evaluation of heterogeneity and developing an understanding of why results vary across studies. Study characteristics such as study type (cohort vs case-control), study quality characteristics (*eg* compliance, blinding *etc*) and population restrictions (*eg RR* for diet as a cause of gastric dilation and volvulus may differ according to the breeds included in a study) should be considered as sources of heterogeneity.

Despite the limitations noted above, methods used in the meta-analysis of RCTs are generally applicable to meta-analyses of observational studies. However, before embarking on such a review, it is important to think about the important ways in which observational studies differ from RCTs and how these will affect a meta-analysis.

28.11.1 Observational studies vs RCTs—bias

As has been noted throughout this text, observational studies are more prone to a variety of biases than RCTs. Consequently, it is particularly important that a meta-analysis not provide a sense of statistical certainty when the contributing studies suffer from serious biases.

In theory, confounding is not possible in a RCT provided that random allocation has been done properly, there is good compliance and the sample size is large (residual confounding from an uneven distribution of a factor across the intervention groups may remain, in small trials). Large sample size in an observational study provides no such guarantee that confounding does not occur. A common confounder may affect all studies being evaluated.

Similarly, RCTs have little opportunity for selection bias because both the treatment and control groups should be representative of the target population provided random allocation has been carried out. The same is not true for observational studies (see Chapter 12).

In RCTs, there should not be any misclassification of exposure (*ie* intervention) although there may be substantial misclassification of the outcome. However, assuming equal follow-up of the 2 study groups, this latter misclassification should be non-differential (see Section 12.6.1) meaning any bias will be toward the null. On the other hand, observational studies may have misclassification of both exposure and outcome and it may be differential or non-differential.

28.11.2 Observational studies vs RCTs—exposure

While the intervention (exposure) may vary across a set of RCTs in some important aspects (*eg* dose, duration of administration), exposure factors are likely to be much more variable across observational studies in which the exposure is not under investigator control. While most RCTs evaluate a single exposure, this limitation does not affect longitudinal and case-control studies which may evaluate a number of possible exposures. While exposure in RCTs is usually known with certainty, exposure level in observational studies is often only recorded in broad categories (*eg* frequency of stall cleaning: daily, more than once a week, weekly *etc*) and these categories may vary across studies. Given that exposure may vary across a wide range, it may be necessary to consider methods for evaluating dose-response evaluations (Dumouchel, 1995).

28.11.3 Observational studies vs RCTs—outcome

In RCTs, the outcome is not as likely to be a rare event (see Section 28.10.6) as it is in an observational study. Consequently, specific methods for dealing with sparse data may be required in a meta-analysis of observational data (Austin *et al*, 1997) (but are beyond the scope of this text).

Because confounding is not likely to be a serious problem in RCTs, it is often not necessary to compute estimates of effect adjusted for potential confounders. In a meta-analysis based on observational studies, this is often very important. If only unadjusted estimates of effect are available, it may be important to adjust these using some form of external adjustment factor.

$$RR_a = RR_u / U \qquad U = RR_u / RR_a \qquad \textbf{\textit{Eq 28.20}}$$

where RR_a=adjusted RR, RR_u=unadjusted RR, and U=measure of confounding bias. U can be estimated from studies which report both the RR_a and the $RR_{u.}$. (For models on a log scale (*eg* logistic model), $U=\beta_u-\beta_a$). For further details, see Chapter 33 of (Rothman *et al*, 2008).

28.12 Meta-analysis of diagnostic tests

Meta-analysis of studies evaluating diagnostic tests is currently an area of considerable research interest and requires some special considerations. A few of the important issues will be identified here, but the reader is referred to other sources for a more detailed coverage of this topic (Egger *et al*, 2001) (Chapter 14) (Devillé *et al*, 2002; Harbord *et al*, 2007; Whiting *et al*, 2003; Whiting *et al*, 2006; Zamora *et al*, 2006).

First, it is important to note that there are many aspects of diagnostic test performance which might be summarised in a meta-analysis. In addition to sensitivity (*Se*) and specificity (*Sp*), one might include likelihood ratios, repeatability, reproducibility, or other measures of test performance (see Chapter 5). Second, most meta-analysis procedures for diagnostic test evaluation studies require that group data (2x2 table cell frequencies) be available in addition to point estimates and SEs (*eg* of *Se* and *Sp*).

In general, if estimates of *Se* and *Sp* are reasonably homogeneous across studies, it may be possible to use standard meta-analysis techniques to compute summary estimates of test performance. For this to be possible, 2 criteria will probably need to be met. First, the test being evaluated should be consistently compared with a good reference test. If studies use a variety of reference tests, it is very likely that there will be considerable variation in the *Se* and *Sp* of the test being evaluated. Second, if the test result is measured on a continuous (or ordinal) scale, a consistent cutpoint needs to be used across all studies. If different cutpoints are used, then there will almost certainly be considerable variation in the *Se* and *Sp* estimates. **Note** Satisfying these 2 criteria certainly does not ensure that study results will be homogeneous.

Given the inverse relationship between *Se* and *Sp*, a summary measure of diagnostic test performance that combines information about both *Se* and *Sp* would be useful. One such measure is the **diagnostic odds ratio** (*DOR*) (Glas *et al*, 2003). It can be computed as:

$$DOR=\frac{TP*TN}{FP*FN}=\frac{\left(\frac{Se}{1-Se}\right)}{\left(\frac{1-Sp}{Sp}\right)}=\frac{LR+}{LR\text{-}}$$

Eq 28.21

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives in a study and *LR*+ and *LR*- are the likelihood ratios of positive and negative test results. It is often necessary to add a small quantity (often 0.5) to each of the 4 values to avoid computational difficulties. The larger the *DOR*, the stronger the diagnostic evidence provided by the test (a value of 1 indicates no diagnostic ability at all). *DOR*s are sometimes pooled in meta-analyses to provide an overall evaluation of the test's capabilities. While the *DOR* has the advantage of combining the *Se* and *Sp* into a single measure, it must be remembered that it does not distinguish between tests with high *Se* and low *Sp* versus low *Se* and high *Sp*.

28.13 Use of meta-analysis

As indicated, the most common use of meta-analysis is for summarising data from a series of controlled trials. They have been used less in veterinary medicine than in human medicine because we seldom have multiple trials of a single product (or closely related group of products) on which to base a meta-analysis. However, with the increasing desire of the profession to have reliable field-based evidence of the efficacy of products used, the availability of clinical-trial data will increase.

Meta-analysis can also be used in research programmes. They might either serve as a ‘definitive study’ by combining the results from many previous studies or they can be used to help design future studies by providing the best estimate of effect for use in sample-size calculations. If a series of studies is being conducted, the results of a meta-analysis can also provide a ‘stopping rule’ by identifying when sufficient evidence of the efficacy of a product exists to warrant halting research on it. In this situation, a cumulative meta-analysis is a useful tool. It shows how the pooled estimate changed as each new study was added. A meta-analysis might also identify factors that strongly influence study results (*ie* contribute to heterogeneity) and guide future research into those effects.

Meta-analysis can also be used to help guide policy decisions. For example, the meta-analysis of the effects of rBST on dairy cattle health and production was one of the pieces of information used by Health Canada when making a decision regarding the registration of the drug for use in Canada (in this case the decision was to not register the drug).

References

Austin H, Perkins LL, Martin DO. Estimating a relative risk across sparse case-control and follow-up studies: a method for meta-analysis Stat Med. 1997; 16: 1005-15.

Baujat B, Mahé C, Pignon J, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials Stat Med. 2002; 21: 2641-52.

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias Biometrics. 1994; 50: 1088-101.

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews J Clin Epidemiol. 2006; 59: 697-703.

DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update Contemp Clin Trials. 2007; 28: 105-14.

DerSimonian R, Laird N. Meta-analysis in clinical trials Control Clin Trials. 1986; 7: 177-88.

Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HCW, van der Windt DAWM, Bezemer PD. Conducting systematic reviews of diagnostic studies: didactic guidelines BMC Med Res Methodol. 2002; 2: 9.

Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis Stat Med. 2006; 25: 2299-322.

Dickersin K. How important is publication bias? A synthesis of available data AIDS Educ Prev. 1997; 9: 15-21.

Dickersin K. Systematic reviews in epidemiology: why are we so far behind? Int J Epidemiol. 2002; 31: 6-12.

Dickersin K, Berlin JA. Meta-analysis: state-of-the-science Epidemiol Rev. 1992; 14: 154-76.

Dohoo I, Stryhn H, Sanchez J. Evaluation of underlying risk as a source of heterogeneity in meta-analyses: a simulation study of Bayesian and frequentist implementations of three models Prev Vet Med. 2007; 81: 38-55.

Dohoo IR, DesCôteaux L, Leslie K, Fredeen A, Shewfelt W, Preston A, Dowling P. A meta-analysis review of the effects of recombinant bovine somatotropin. 2. Effects on animal health, reproductive performance, and culling Can J Vet Res. 2003a; 67: 252-64.

Dohoo IR, Leslie K, DesCôteaux L, Fredeen A, Dowling P, Preston A, Shewfelt W. A meta-analysis review of the effects of recombinant bovine somatotropin. 1. Methodology and effects on production Can J Vet Res. 2003b; 67: 241-51.

Duffield TF, Rabiee AR, Lean IJ. A meta-analysis of the impact of monensin in lactating dairy cattle. Part 1. Metabolic effects J Dairy Sci. 2008; 91: 1334-46.

Dumouchel W. Meta-analysis for dose-response models Stat Med. 1995; 14: 679-85.

Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis Biometrics. 2000; 56: 455-63.

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test BMJ. 1997; 315: 629-34.

Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies BMJ. 1998; 316: 140-4.

Egger M, Smith G, Altman D. Systematic Reviews in Health Care. Meta-analysis in context. BMJ Books; London. 2001.

Fourichon C, Seegers H, Malher X. Effect of disease on reproduction in the dairy cow: a meta-analysis Theriogenology. 2000; 53: 1729-59.

Freemantle N, Cleland J, Young P, Mason J, Harrison J. beta Blockade after myocardial infarction: systematic review and meta regression analysis BMJ. 1999; 318: 1730-7.

Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results J Clin Epidemiol. 2006; 59: 7-10.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance J Clin Epidemiol. 2003; 56: 1129-35.

Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews Stat Med. 2002; 21: 1503-11.

Greenland S. Invited commentary: a critical look at some popular meta-analytic methods Am J Epidemiol. 1994; 140: 290-6.

Harbord R, Higgins J. Meta-regression in Stata The Stata Journal. 2008; 8: 493-519.

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies Biostatistics. 2007; 8: 239-51.

Hedges LV, Pigott TD. The power of statistical tests in meta-analysis Psychol Methods. 2001; 6: 203-17.

Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned J Clin Epidemiol. 2006; 59: 1249-56.

Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1 [updated September 2008]. : The Cochrane Collaboration, Available from www.cochrane-handbook.org. 2008.

Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice J Health Serv Res Policy. 2002; 7: 51-61.

Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis Stat Med. 2002; 21: 1539-58.

Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression Stat Med. 2004; 23: 1663-82.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses BMJ. 2003; 327: 557-60.

Higgins JPT, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: Combining results reported on log-transformed or raw scales Stat Med. 2008; 27: 6072-92.

Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials Stat Med. 2002; 21: 1625-34.

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions Cochrane Database Syst Rev. 2007; 2: Art. No.: MR000010.

Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis Stat Med. 2006; 25: 2688-99.

Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials BMJ. 2001; 323: 42-6.

Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate Stat Med. 2003; 22: 2693-710.

Petitti DB. Approaches to heterogeneity in meta-analysis Stat Med. 2001; 20: 3625-33.

Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed Int J Epidemiol. 2002; 31: 150-3.

Rothman K, Greenland S, Lash T. Modern Epidemiology, 3rd Ed. Lippincott Williams & Wilkins; Philadelphia. 2008.

Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells Stat Med. 2009; 28: 721-38.

Sanchez J, Dohoo I, Carrier J, DesCôteaux L. A meta-analysis of the milk-production response after anthelmintic treatment in naturally infected adult dairy cows Prev Vet Med. 2004; 63: 237-56.

Sánchez-Meca J, Marín-Martínez F, Chacón-Moscoso S. Effect-size indices for dichotomized outcomes in meta-analysis Psychol Methods. 2003; 8: 448-67.

Sargeant JM, Rajic A, Read S, Ohlsson A. The process of systematic review and its application in agri-food public-health Prev Vet Med. 2006; 75: 141-51.

Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials Stat Med. 1998; 17: 1923-42.

Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies Stat Med. 2007; 26: 1964-81.

Song F. Exploring heterogeneity in meta-analysis: is the L'Abbé plot useful? J Clin Epidemiol. 1999; 52: 725-30.

Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature J Clin Epidemiol. 2000; 53: 1119-29.

Surgeon General. Smoking and Health. 1964. US Government Printing Office. Washington DC.

Sutton AJ, Higgins JPT. Recent developments in meta-analysis Stat Med. 2008; 27: 625-50.

Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data Stat Med. 2004; 23: 1351-75.

Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias J Clin Epidemiol. 2005; 58: 894-901.

Thiessen Philbrook H, Barrowman N, Garg AX. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: a case study of changes in renal function after living kidney donation J Clin Epidemiol. 2007; 60: 228-40.

Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002; 21: 1559-73.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies Epidemiology. 2007; 18: 800-4.

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews BMC Med Res Methodol. 2003; 3: 25.

Whiting PF, Westwood ME, Rutjes AWS, Reitsma JB, Bossuyt PNM, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies BMC Med Res Methodol. 2006; 6: 9.

Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, Barrowman NJ. A systematic review identifies a lack of standardization in methods for handling missing variance data J Clin Epidemiol. 2006; 59: 342-53.

Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data BMC Med Res Methodol. 2006; 6: 31.