ABSTRACT

The objectives of this paper are to provide an introduction to meta-analysis and systematic review and to discuss the rationale for this type of research and other general considerations. We highlight methods used to produce a rigorous meta-analysis and discuss some aspects of interpretation of meta-analysis drawing on examples from the animal and veterinary science literature. Meta-analysis is a rapidly expanding area of research that has been relatively underutilized in animal and veterinary science. It is a quantitative, formal, epidemiological study design used to systematically assess previous research studies to derive conclusions about that body of research. Outcomes from a meta-analysis may include a more precise estimate of the effect of treatment or risk factor for disease, or other outcomes, than any individual study contributing to the pooled analysis. The examination of variability or heterogeneity in study results is also a critical outcome. The benefits of meta-analysis include a consolidated and quantitative review of a large, and often complex, sometimes apparently conflicting, body of literature. Meta-analytic methods place less emphasis on dichotomous outcomes from null hypothesis significance testing and greater emphasis on determining the magnitude and the precision of an effect of interest. A substantial benefit of meta-analysis is the potential to investigate new hypotheses using existing data, both through the development of a priori hypotheses and by examination of the heterogeneity in study responses. The specification of the outcome and hypotheses that are tested is critical to the conduct of meta-analyses, as is a sensitive literature search. A failure to identify the majority of existing studies can lead to erroneous conclusions; however, there are methods of examining data to identify the potential for studies to be missing; for example, by the use of funnel plots. Many of the statistical methods to conduct meta-analysis are widely used. Bayesian methods are well suited to meta-analysis. The post-hoc methods used to evaluate heterogeneity and publication bias, which include the $I^2$ statistic, L’Abbé plots, Galbraith plots, Rosenthal’s N, and influential study analysis are exclusively used in meta-analysis. Examples where meta-analyses have been repeated in animal science or veterinary medicine show good consistency in estimates of effect. Findings of studies to date have provided new understandings of rumen modifiers, milk fever, parasite control, mastitis, somatotropin, and reproductive manipulations. Rigorously conducted meta-analyses are useful tools to improve animal well-being and productivity. The need to integrate findings from many studies ensures that meta-analytic research is desirable and the large body of research now generated makes the conduct of this research feasible.

Key words: meta-analysis, dairy cow, sample size, case definition

INTRODUCTION

The scientific investigation of animal health and reproduction seeks to provide findings that can be confidently applied to improve health, production, and reproduction. The classical cycle of observation providing a hypothesis, testing of the hypothesis, and refinement of the hypothesis by further testing can now be extended to include the investigation of hypotheses through production of estimates of effect and confidence in these effects using a series of existing studies. The objectives of this paper are to introduce meta-analysis and systematic review, and to discuss the rationale for this type of research and other general considerations. We highlight the methods used to produce a rigorous meta-analysis and discuss some aspects of interpretation of meta-analysis drawing on examples from the animal and veterinary science literature.
Meta-Analysis and Systematic Review

Glass (1976) defined meta-analysis as “The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.” Meta-analysis is a quantitative, formal, epidemiological study design used to systematically assess the results of previous research to derive conclusions about that body of research. Typically, but not necessarily, the study is based on randomized, controlled clinical trials. Outcomes from a meta-analysis may include a more precise estimate of the effect of treatment or risk factor for disease, or other outcomes, than any individual study contributing to the pooled analysis. Identifying sources of variation in responses; that is, examining heterogeneity of a group of studies, and generalizability of responses can lead to more effective treatments or modifications of management. Examination of heterogeneity is perhaps the most important task in meta-analysis. The Cochrane collaboration (Cochrane Collaboration, 2008) has been a long-standing, rigorous, and innovative leader in developing methods in the field. Major contributions include the development of protocols that provide structure for literature search methods, and new and extended analytic and diagnostic methods for evaluating the output of meta-analyses. Use of the methods outlined in the handbook should provide a consistent approach to the conduct of meta-analysis.

Meta-analyses are a subset of systematic review. A systematic review attempts to collate empirical evidence that fits prespecified eligibility criteria to answer a specific research question (Sargeant et al., 2006). The key characteristics of a systematic review are a clearly stated set of objectives with predefined eligibility criteria for studies; an explicit, reproducible methodology; a systematic search that attempts to identify all studies that meet the eligibility criteria; an assessment of the validity of the findings of the included studies (e.g., through the assessment of risk of bias); and a systematic presentation and synthesis of the attributes and findings from the studies used. Systematic methods are used to minimize bias, thus providing more reliable findings from which conclusions can be drawn and decisions made than traditional review methods (Antman et al., 1992; Oxman and Guyatt, 1993). Systematic reviews need not contain a meta-analysis—there are times when it is not appropriate or possible; however, many systematic reviews contain meta-analyses. Westwood et al. (1998a,b) provided a quantitative, systematic review of protein nutrition in cattle and included several pooled effect estimates and meta-analyses. These provided evidence of the strength of correlation between nitrogen concentrations in the rumen, blood, and milk and the effects on first-service conception rate of increasing CP in the diet and effects of increasing protein protection with isonitrogenous diets.

The benefits of meta-analysis include a consolidated and quantitative review of a large, and often complex, sometimes apparently conflicting, body of literature; for example, the effects of recombinant somatotropin (rbST) on milk production and health of cattle (Dohoo et al., 2003a,b). Meta-analysis can be used by regulatory authorities to provide a quantitative evaluation of the efficacy of a product available for registration. The hypotheses that can be raised in a meta-analysis may extend well beyond those examined in the original studies; for example, meta-analysis was used to evaluate the multivariable effects of macromineral concentrations in the diet on the risk of milk fever (Oetzel, 1991), whereas most of the original studies were univariable in design.

The value of meta-analysis to address the marked increase in literature available to scientists and practitioners in most fields of scientific endeavor has been recognized. These methods have extended beyond the social sciences and human medicine, where they were first developed and adopted, and span disciplines from astronomy to zoology (Petticrew, 2001). From a few early publications, the number of published meta-analyses in medical science had increased to 400 per year in the year 2000 (Lee et al., 2001). While researching this paper, we identified more than 1,000 papers published using quantitative or systematic review in medical science in the year 2008. Statistical techniques to combine results from separate randomized controlled trials, economic evaluations, and epidemiological studies have become more common in animal and veterinary science, possibly because of the increasing size of the research literature, and partly because of the demand for rigorous efficacy and safety standards. Despite a wealth of suitable studies, however, there has not been growth in meta-analysis in animal and veterinary science comparable to that in medicine. Our literature search using PubMed, CAB (Commonwealth Agricultural Bureaux), and Scirus identified only 150 formal meta-analyses and quantitative reviews in cattle. Early examples of meta-analysis in animal and veterinary science include Oetzel (1991), Enevoldsen (1993) investigating factors influencing milk fever, Morgan and Lean (1993) examining effects of GnRH on the probability of pregnancy, and Gross et al. (1999), who examined effects of anthelmintic treatment on milk production. Some areas of veterinary and animal science have benefited from well-executed quantitative review including treatment of parasites (Gross et al., 1999; Sanchez et al., 2004), disease conditions (Fourichon et al., 2000),
rumen modification (Lean and Wade, 1997; Duffield et al., 2008a,b,c), rbST (Dohoo et al., 2003a,b), and prediction of milk fatty acids (Moate et al., 2008).

The inclusion of observational medical studies in meta-analyses led to considerable debate over the validity of meta-analytical approaches, as there was necessarily a concern that the observational studies were likely to be subject to unidentified sources of confounding and risk modification (Greenland 1994a,b; Olkin, 1994). Pooling such findings may not lead to more certain outcomes. Animal health and production has a distinct advantage over medical research in being more able to make use of randomized controlled clinical trials.

Meta-analyses are conducted to assess the strength of evidence present on a disease, treatment, or aspects of production or reproduction. One aim is to determine whether an effect exists; another aim is to determine whether the effect is positive or negative and, ideally, to obtain a single summary estimate of the effect. The results of a meta-analysis can improve precision of estimates of effect, answer questions not posed by the individual studies, settle controversies arising from apparently conflicting studies, and generate new hypotheses. In particular, the examination of heterogeneity is vital to the development of new hypotheses. Although dichotomous outcomes are inherent to null hypothesis testing, Nakagawa and Cuthill (2007) note that a key component of meta-analytic thinking is to move away from desiring dichotomous outcomes from null hypothesis significance testing and to place greater emphasis on determining the magnitude of an effect of interest and the precision of estimate of that magnitude of effect. The principles of meta-analysis mean that the cases in one study are not directly compared with those in another, and each study is analyzed separately, unless dealing with pooled analyses using individual data from original studies.

A large number of meta-analyses are undertaken with the broad aim of summarizing existing evidence on a subject; for example, Gross et al. (1999) to evaluate the effect of anthelmintic treatment on milk production in cattle and Robert et al. (2006) to evaluate mastitis and the effects of antibiotic treatment of dry cows. Others are undertaken to inform specific decisions (Sanchez et al., 2004) to evaluate the effect of different classes of anthelmintics on milk production of cattle and may be incorporated in economic models in a decision analysis framework. A well-conducted meta-analysis can provide unbiased overviews of the available evidence on product claims. There is considerable interest in the use of meta-analysis to evaluate the use of new technologies for regulatory evaluation purposes; for example, to assess efficacy and safety of new therapeutics in animal health, production, and reproduction.

**GENERAL CONSIDERATIONS IN META-ANALYSIS**

**Individual or Aggregated Data**

The majority of meta-analyses are based on a series of studies to produce a point estimate of an effect and measures of the precision of that estimate. However, methods have been developed (Stewart and Clarke, 1995; Simmonds et al., 2005) for the meta-analyses to be conducted on data obtained from original trials. This approach may be considered the “gold standard” in meta-analysis because it offers advantages over analyses using aggregated data, including a greater ability to validate the quality of data and to conduct appropriate statistical analysis. Further, it is easier to explore differences in effect across subgroups within the study population than with aggregated data. This approach relies on cooperation between researchers who conducted the relevant studies. We used this approach to examine the effect of monensin on time to conception (Duffield et al., 2008c) using raw data obtained from those involved in the original studies. Researchers who are aware of the potential to contribute or conduct these studies will provide and obtain additional benefits by careful maintenance of original databases and making these available for future studies.

**Study Power**

A major rationale for use of meta-analysis is to increase the statistical power of hypotheses. Most disease and many reproductive outcomes are dichotomous and, consequently, require large numbers of animals to determine statistically significant differences between treatments. Figure 1 provides estimates of numbers of animals required to determine differences between 2 groups of subjects with statistical power of 0.9 and \( \alpha = 0.05 \) for some common diseases at anticipated incidences. Similarly, to detect a 5% difference in conception rates to a single service with an \( \alpha = 0.05 \) and statistical power of 0.9 requires 460 cows per group.

Very few of the studies of the effects of GnRH on the probability of pregnancy assessed by Morgan and Lean (1993) had a statistical power of 0.9; the average study size was 238 (ranged from 10 to 1,797) cows per group. A common way of presenting data in a meta-analysis is a forest plot (Figure 2). These data, extracted from Morgan and Lean (1993), show that very few of the studies had confidence intervals that did not include 1, indicating that these studies did not determine a statistically significant difference in conception. Consequently, most studies concluded that there was “no effect” of GnRH administered at the time of insemination, and reviews of the effect concluded that treatment (with GnRH)
had no effect on conception (Wright and Malmo, 1992). Figure 2, however, shows that frequently the point effect in studies was large, albeit not significant, especially in repeat breeder cows. Morgan and Lean (1993) found that GnRH treatment was effective in increasing the probability of pregnancy in repeat breeder cattle (risk ratio = 1.24, 95% confidence interval: 1.11 to 1.38), although the effect was heterogeneous.

Similarly, the large number of observations available at the time of analysis on the effect of monensin on first-service conception rate lead Lean and Wade (1997) to conclude that “These apparently favorable physiological responses to monensin in heifers have not been reflected in reproductive performance in adult, lactating dairy cattle. A meta-analysis of first-service conception rates shows that there was no significant effect of monensin treatment on first-service conception rates. A large number of studies with either a positive or negative response to treatment would be required to alter this result and the results were homogenous (P > 0.25) indicating that there was consistency of response to treatment.” This conclusion was validated by Duf- field et al. (2008c) using many more cattle across many different production systems in 9 countries.

In both cases, an increased precision of estimate of effect over any single, even very large, study and increased external validity of the meta-analysis that used data from a large number of herds and production systems, provided valuable information on the effects of treatment.

Evidence in Animal Studies

There are many examples of observational epide- miological study designs that provide useful evidence of disease or reproduction that include case-control studies (Davis et al., 1980), and cohort studies (Curtis et al., 1984; Degaris et al., 2008). We recommend, in general, that the evidential base used in a meta-analysis consist of randomized controlled clinical studies or prospective cohort studies as these should provide the most rigorous data. Studies on the effects of disease on reproductive performance of dairy cattle (Fourichon et al., 2000) and effects of milk production on disease and reproduction in dairy cattle (Ingvartsen et al., 2003) are examples, however, of the appropriate application of observational studies in quantitative review. Figure 3 shows that there is a range of different study types that can be used to examine health and reproduction and that these provide different insights into these processes. Figure 3 also shows that the studies vary in their validity, with larger multi-site, multicenter studies having great external validity and smaller, more tightly controlled, laboratory or experimental station studies having greater internal validity.

One benefit of meta-analysis is that findings should have a higher external validity than any single study. Evidence obtained from studies varies in value. Smaller studies, in general, have less external validity than larger studies; that is, the results tend to be quite specific to the population under study and care should be taken extrapolating these to outside populations. However, such smaller studies are generally more amenable to tight control of factors that could confound the observations made. Consequently, these smaller studies have a high level of internal validity and we should be quite confident that the results described reflect the effect of perturbation of the system, provided that the studies are well conducted. As noted previously, large studies are typically required to examine the effects of treatment on animal health with sufficient statistical power. These studies will have greater external validity; that is, be more likely to reflect responses obtained in the field. However, these larger studies may be more susceptible to measurement errors and potential confounding, and therefore have lower internal validity than smaller studies. In statistical analysis, weighting of studies reflects differences in the value of evidence provided and this will reflect both the size of the study and the variance of measurements.

Case Definition and Implications

The definition of the outcome (i.e., health disorder or production measure) is critical to the conduct of successful meta-analyses. A failure to adequately define the condition or to identify a lack of consistency in the
Figure 2. Risk ratios (RR) [and their 95% confidence interval (CI) and weights for individual trials] determined from the results of 42 trials comparing risk of pregnancy (PR) conception at 1st and 2nd insemination and overall risk of pregnancy in dairy cows injected with GnRH or analog at the time of insemination and repeat breeder cows (data obtained from Morgan and Lean, 1993). Box sizes are proportional to the inverse variance of the estimates. Summary estimates of treatment effects are shown using: 1) a fixed effects approach (Mantel-Haenszel = M-H subtotal), 2) a random effects approach (DerSimonian and Laird = D+L subtotal), and 3) the estimated predicted interval of a future trial; this prediction can provide further evidence of the effectiveness of treatment in future studies. A full list of the studies referenced here is available online (http://jds.fass.org/content/vol92/issue8/).
definition in studies can lead to problems of interpretation in meta-analysis.

Definitions of disease often reflect historical understandings of disease (e.g., “milk fever”) or terms that reflect aspects of the perceived etiology (e.g., “grass tetany”). Classification of disease should be accurate and reflect standardized and repeatable criteria. Attempts have been made to standardize reproductive terminology; for example, Hubbert et al. (1972). These efforts have provided some clarity to what was an unclear literature. However, there have been few consolidated attempts to define disease processes. Kelton et al. (1998) used an approach of providing standardized definitions of disease. Further, in the case of disease, there needs to be a detailed examination of Evans’ postulates (Evans, 1976) to support the definition. In the case of metabolic disease, we propose the following postulates consistent with those of Evans: 1) the definition of disease should be consistent with current understandings of the biochemical basis of the disorder or reflect modifications thereof; 2) the proportion of individuals with the disease should be significantly higher in those exposed to the supposed cause than in those who are not; 3) exposure to a supposed cause should be present more commonly in those with than in those without the disease, when all other risk factors are held constant; 4) the number of new cases of disease should be significantly higher in those exposed to a supposed cause than in those not so exposed, as shown by prospective studies; 5) a spectrum of host responses from mild to severe should follow exposure to a supposed cause along a logical biological gradient; 6) a measurable change in metabolism should appear regularly following exposure to a supposed cause in those lacking this response before exposure, or should increase in magnitude if present before exposure; this pattern should not occur in individuals not so exposed; 7) experimental reproduction of the disease should occur with greater frequency in animals or man appropriately exposed to a supposed cause than in those not so exposed; 8) elimination (e.g., removal of a specific risk factor) or modification (e.g., alteration of a deficient diet) of the supposed cause should decrease the frequency of occurrence of the disease; and 9) all relationships and associations should be biologically and epidemiologically credible.

Approaches consistent with these postulates have been used to define the metabolic diseases ketosis (Lean et al., 1994; Duffield et al., 2009) and acidosis (Bramley et al., 2008). Curtis (1997) and Sheldon et al. (2006) also used careful definitions of uterine infection and examined these definitions using risk factors for and outcomes of uterine infection. Importantly, all these re-
searchers demonstrated that the definitions developed either predicted loss of production or increased risk of other disease.

Furthermore, it is not always simple to make decisions about the appropriateness of the grouping of a series of studies. Rabiee et al. (2005) decided to include the Pre-synch and Co-synch methods, and minor modifications of these for synchronizing estrus in category called “modified Ovsynch.” Advantages of that decision included the opportunity to examine the most recent developments in the field; hence, those methods of ovulation synchrony with the least data at the time of investigation. It was also considered likely that these treatments provided a consistent physiological approach to the synchrony of ovulation. A critical validation of the determination to use these studies as a single category was the post hoc finding of strong homogeneity of the group of studies, supporting the hypothesis that the modified Ovsynch group of treatments was consistent in physiological action. It is vital to stress that meta-analysis, or for that matter, any other study design, needs to consider, in detail and with rigor, the biological basis of the study.

The definition of disease is critical to the process of investigation. If there is a nondifferential error in the definition of disease (i.e., errors of classification are not related to potential risk factors investigated), the effect is to bias the result toward the null. If there are differential errors in definition, the results are unpredictable and can drive the findings toward the null or toward a positive but potentially spurious finding (Kelsey et al., 1996).

**METHODS**

**Literature Search**

A sound meta-analysis is characterized by a thorough and disciplined literature search. A clear definition of hypotheses to be investigated provides the framework for such an investigation. It is important to obtain all relevant studies, because loss of studies can lead to bias in the study. Typically, published papers and abstracts are identified by a computerized literature search of electronic databases that can include Commonwealth Agricultural Bureau (CAB), Biological Abstracts, Life Sciences, ScienceDirect (www.sciencedirect.com), Scirus (www.scirus.com/srsapp), PubMed (www.ncbi.nlm.nih.gov/entrez/query.fcgi), Agricola (agricola.nal.usda.gov), ISI Web of Knowledge (http://www.isiwebofknowledge.com), and Google Scholar (http://scholar.google.com). Database searches should be augmented with hand searches of library resources for relevant papers, books, abstracts, and conference proceedings. Cross-checking of references, citations in review papers, and communication with scientists who have been working in the relevant field are important methods used to provide a comprehensive search. Communication with pharmaceutical companies manufacturing and distributing test products can be appropriate for studies examining the use of pharmaceutical interventions.

It is not feasible to find absolutely every relevant study on a subject. Some or even many studies may not be published, and those that are might not be indexed in computer-searchable databases. The reviewers should attempt to be sensitive; that is, find as many studies as possible, to minimize bias and be efficient. It may be appropriate to frame a hypothesis that considers the time over which a study is conducted or to target a particular subpopulation. The decision whether to include unpublished studies is difficult. Although language of publication can provide a difficulty, it is important to overcome this difficulty, provided that the populations studied are relevant to the hypothesis being tested. Sanchez et al. (2004), for example, sought to include papers in English, Spanish, French, Portuguese, and Italian, and their search included peer-reviewed journals, abstracts, conference proceedings, and theses.

**Inclusion or Exclusion Criteria and Potential for Bias**

Studies are chosen for meta-analysis based on inclusion criteria. If there is more than one hypothesis to be tested, separate selection criteria should be defined for each hypothesis. Inclusion criteria are ideally defined at the stage of initial development of the study protocol. The rationale for the criteria for study selection used should be clearly stated.

One important potential source of bias in meta-analysis is the loss of trials and subjects. Ideally, all randomized subjects in all studies satisfy all of the trial selection criteria, comply with all the trial procedures, and provide complete data. Under these conditions, an “intention-to-treat” analysis is straightforward to implement; that is, statistical analysis is conducted on all animals that are enrolled in a study rather than those that complete all stages of study considered desirable. However, not all animals provide complete data in large field studies. We strongly recommend that published studies provide full details of subject loss and recommend that, whenever possible, papers are analyzed on an intention-to-treat basis.

Further, not all studies are completed, because of protocol failure, treatment failure, or other factors. Nonetheless, missing subjects and studies can provide important evidence. It is desirable to obtain data from all relevant randomized trials, so that the most appro-
Appropriate analysis can be undertaken. Chan et al. (2004) and Dwan et al. (2008) have discussed the significance of missing trials to the interpretation of intervention studies in medicine. Journal editors and reviewers need to be aware of the existing bias toward publishing positive findings and ensure that papers that publish negative or even failed trials be published, as long as these meet the quality guidelines for publication.

There are occasions when authors of the selected papers have chosen different outcome criteria for their main analysis. In practice, it may be necessary to revise the inclusion criteria for a meta-analysis after reviewing all of the studies found through the search strategy. Variation in studies reflects the type of study design used, and type and application of experimental and control therapies, whether or not the study was published, and, if published, subjected to peer review, and the definition used for the outcome of interest. There are no standardized criteria for inclusion of studies in meta-analysis. Universal criteria are not appropriate, however, because meta-analysis can be applied to a broad spectrum of topics. Published data in journal papers should also be cross-checked with conference papers to avoid repetition in presented data.

Clearly, unpublished studies are not found by searching the literature. It is possible that published studies are systemically different from unpublished studies; for example, positive trial findings may be more likely to be published. Therefore, a meta-analysis based on literature search results alone may lead to publication bias. Efforts to minimize this potential bias include working from the references in published studies, searching computerized databases of unpublished material, and investigating other sources of information including conference proceedings and graduate dissertations.

Quality scores have been used to include or exclude studies from a meta-analysis (Jüni et al., 2001), and ranking methods were developed primarily for studies that sought to include data provided from observational, as opposed to experimental studies, and can be subjective. Before assessing study quality, a quality assessment protocol and data forms should be developed. The goal of this process is to reduce the risk of bias in the estimate of effect. To reduce bias toward a particular view, reviewers should read only the methods and results of the published reports, and all identifying information, such as details providing the journal of publication, authors, institution in which the authors work, are removed from copies of reports given to assessors. The study design, including details of the method of randomization of subjects to treatment groups, criteria for eligibility in the study, blinding, method of assessing the outcome, and handling of protocol deviations are important features defining study quality. When studies are excluded from a meta-analysis, reasons for exclusion should be provided for each excluded study. Usually, more than one assessor decides independently which studies to include or exclude, together with a well-defined checklist and a procedure that is followed when the assessors disagree. Two people familiar with the study topic perform the quality assessment for each study. This is followed by a consensus meeting to discuss the studies excluded or included. Practically, the blinding of reviewers from details of a study such as authorship and journal source is difficult.

We do not generally recommend the use of quality scores to exclude studies. We have examined the effect of including or excluding studies based on “quality” criteria in several data sets and found little difference in outcomes of the pooled assessments. We consider that the post hoc evaluation of the studies objectively ranked on quality attributes, however, can be of value in understanding sources of heterogeneity in a group of studies.

Statistical Analysis

The most common measures of effect used for dichotomous data are the risk ratio (also called relative risk) and the odds ratio. The dominant method used for continuous data are standardized mean difference (SMD) estimation. Methods used in meta-analysis for post hoc analysis of findings are relatively specific to meta-analysis and include heterogeneity analysis, sensitivity analysis, and evaluation of publication bias.

All methods used should allow for the weighting of studies. The concept of weighting reflects the value of the evidence of any particular study. For dichotomous outcomes, the weighting for each study is a function of the number of animals enrolled in the study and in the proportion with the outcome of interest. For an example, see equations 1 to 3. In effect size estimates, the weighting reflects the number of animals used and the variance of the study. Usually, studies are weighted according to the inverse of their variance (Egger et al., 2001). It is important to recognize that smaller studies, therefore, usually contribute less to the estimates of overall effect. This concept is demonstrated visually in Figure 2, in which the size of the box in the forest plot indicates the contribution of the study to the overall effect. However, well-conducted studies with tight control of measurement variation and sources of confounding contribute more to estimates of overall effect than a study of identical size less well conducted.

The following texts, Web site, and papers should be consulted for greater detail on the statistical methods used (Cochrane Collaboration, 2008; http://www.cochrane-handbook.org/; Petiti, 1994; Stangl and Berry,
2000; Egger et al., 2001; Whitehead, 2002). The computer program Stata (Intercooled Stata V.10.2, StataCorp., College Station, TX) provides a comprehensive suite of programs that can be used in meta-analysis. Some detail on the common statistical methods used is presented and examples of when these have been used in studies using cattle are provided.

It has been argued that Bayesian models are more appropriate to decision making in medical policy than frequentist approaches (Spiegelhalter et al., 2000; Congdon, 2001). Furthermore, some workers have suggested that meta-analyses are a “natural” fit for Bayesian statistical approaches (Lindley, 1972). Examples of Bayesian analysis are rare in the animal and veterinary science literature. Rabiee et al. (2004, 2005) chose to use a random-effects, Bayesian model to examine the use of controlled release intravaginal devices that release progesterone and Ovsynch programs on reproductive performance of dairy cows and outlined the statistical methods used. The advantage perceived for this analysis was a benefit of Bayesian approaches in which data, particularly the number of studies, were sparse and the outcomes are more robust than frequentist approaches (Stangl and Berry, 2000). Other advantages of the Bayesian approaches include the production of credible (confidence) intervals rather than P-values, that all parameters are random rather than fixed effects, and the potential to incorporate existing knowledge or even subjective estimates of effect in “informed priors” (Stangl and Berry, 2000; Congdon, 2001; Whitehead, 2002).

One of the foremost decisions to be made when conducting a meta-analysis is whether to use a fixed-effects or a random-effects model. A fixed-effects model is based on the assumption that the sole source of variation in observed outcomes is that occurring within the study; that is, the effect expected from each study is the same. Consequently, it is assumed that the models are homogeneous; there are no differences in the underlying study population, no differences in subject selection criteria, and treatments are applied the same way (Stangl and Berry, 2000). Random-effects models have an underlying assumption that a distribution of effects exists, resulting in heterogeneity among study results. Consequently, as software has improved, random-effects models and Bayesian approaches that require greater computing power have become more frequently conducted. This is desirable because the strong assumption that the effect of interest is the same in all studies is frequently untenable. Whitehead (2002) recommends comparing the fixed effects and random effect models developed as this process can yield insights to the data.

Multilevel modeling is now an accepted statistical analysis method for hierarchical data (Goldstein, 1995). Meta-analysis can also be viewed as a special case of multilevel analysis (Hox and de Leeuw, 2002). When we have a hierarchical data set, with subject within studies at the first level and studies at the second level, the multilevel approach is appropriate. If the original data are available, a standard multilevel analysis can be carried out, predicting the outcome variable using the available individual and study-level explanatory variables. Access to the original raw data is unusual; more frequently, the published results are provided in the form of P-values, means, standard deviations, or correlation coefficients. Raudenbush and Bryk (1985), Bryk and Raudenbush (1992), Kalaian and Raudenbush (1996), and van den Noortgate and Onghena (2003) showed how the multilevel approach could be applied in meta-analysis. Kalaian and Raudenbush (1996) also demonstrated the use of a mixed model for meta-analysis as a special case of the multilevel regression model. The analysis is performed on available statistics instead of raw data, and as a result, some specific restrictions must be imposed on the model (Hox and de Leeuw, 2002). The major advantage of using multilevel analysis instead of classical meta-analysis methods is flexibility. In multilevel meta-analysis, it is simple to include study characteristics as explanatory variables in the model. If there is a hypothesis that the study characteristics can influence the outcomes, these will be included on a priori grounds in the analysis. Alternatively, when it is found that the study outcomes are heterogeneous, the available study variables can be explored to explain the heterogeneity.

Hierarchical models or multilevel models for meta-analysis require the application of some specialized approaches (Thompson et al., 2001). This is because the information in a meta-analysis is usually derived from 2 levels: studies at the higher level and participants within studies at the lower level. Sometimes additional levels of data may be relevant; for example, centers in a multicenter trial or clusters in a cluster-randomized trial. A hierarchical framework is appropriate whether meta-analysis is conducted on summary statistic information or individual patient data (Turner et al., 2000). Such a framework is particularly relevant when random effects are used to represent unexplained variation in effect estimates among studies.

Hierarchical models are useful in several contexts and they can be used to 1) allow for the imprecision of the variance estimates of treatment effects within studies; 2) allow for the imprecision in the estimated between-study variance estimate, 3) provide methods that explicitly model binary outcome data, rather than
Dichotomous Data. Given that many of the earlier meta-analyses related to cattle were on the effect of reproductive manipulations on conception or pregnancy, a risk ratio or relative risk (RR) analysis was used. This method used data extracted from individual studies to calculate the RR for individual studies and subsequently the Mantel-Haenszel (MH) test, a fixed-effect model, was used to calculate pooled effects.

The MH method weights studies by sample size and allows stratification by study. Separate 2 × 2 tables are constructed and a pooled estimate of relative risk calculated. A chi-square statistic with 1 df is used to compute a confidence interval around the relative risk. The following formulae and methods are used for the calculation of MH relative risk (Rothman, 1986):

\[ \text{Mantel-Haenszel relative risk (RR)} = \frac{\sum (a_i N_{0i} / T_i)}{\sum (b_i N_{1i} / T_i)}, \]

where \( a, b, N_0, N_1, \) and \( T_i \) are calculated from 2 × 2 tables constructed for each trial (Table 1).

Variance of the \( \text{RR}_{\text{MH}} \) is calculated with the following formula:

\[ \text{Var} \ \ln(RR_{\text{MH}}) = \frac{\sum_{i=1}^{N} \frac{M_{1i} N_{0i} N_{0i}}{T_i^2}}{\left( \sum_{i=1}^{N} \frac{b_i N_{1i}}{T_i} \right)^2}, \]

where \( a, b, N_0, N_1, M_1, \) and \( T \) are calculated from 2 × 2 tables constructed for each trial and \( \ln \) is the natural log. Confidence intervals for the \( \text{RR}_{\text{MH}} \) can then be calculated using this variance:

\[ 95\% \ CI = \exp[\ln(RR_{\text{MH}}) + 1.96 \sqrt{\text{var} [\ln(RR_{\text{MH}})]}]. \]

<table>
<thead>
<tr>
<th>Table 1. Construction of a 2 × 2 table for the calculation of Mantel-Haenszel relative risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
</tr>
<tr>
<td>:--------</td>
</tr>
<tr>
<td>Cases</td>
</tr>
<tr>
<td>Non-cases</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
control group standard deviation. An ES so derived is a “Z” or normal probability score, which reflects the probability that the treated group mean arose from the control group population (Hedges and Olkin, 1985). Effect size estimates are pooled using the method of Hedges (1982). Differences between treated and control group means are weighted by sample size and averaged to provide an indication on the extent of the pooled treatment effect.

**ES Calculation.** In the fixed-effects case, all the ES estimate the single population effect size. If each ES could be measured with an infinite sample, all the ES in the meta-analysis would be the same.

The ES for each study is calculated using the following formula

\[
ES = \frac{\bar{X}_e - \bar{X}_c}{S_p},
\]

where \(\bar{X}_e\) is the experimental group mean, \(\bar{X}_c\) is the control group mean, and \(S_p\) is the pooled standard deviation. The pooled standard deviation of the control and treatment groups is used as the estimate of \(S_p\) as described by Mullen and Miller (1991). Standard deviation \((S)\) is extracted from the papers reporting data as follows:

Where an overall standard error of the mean \((SEM_p)\) is reported for the treatment and control groups, \(S_p\) is calculated thus:

\[
S_p = SEM_p \times \sqrt{n_p},
\]

where \(n_p\) is the total number of cows in the treatment and control groups (Huntsberger and Billingsley, 1987).

Where the SEM is reported for the treatment and control groups separately, the standard deviation is calculated with the formula above for each sample group and the standard deviations are then pooled.

Where papers reported standard deviations for the treatment and control groups, these are also pooled with the formula:

\[
S_p = \sqrt{\frac{(n_e - 1)(s_e)^2 + (n_c - 1)(s_c)^2}{n_e + n_c - 2}},
\]

where \(n_e\) and \(n_c\) are the number of cows in the experimental and control groups, respectively, and \(s_e\) and \(s_c\) are the standard deviations for the experimental and control groups (Hedges and Olkin, 1985).

The overall ES for all trials \((ES_o)\) is calculated weighted for sample size of each trial:

\[
ES_o = \frac{\sum_{i=1}^{k} n_i ES_i}{\sum_{i=1}^{k} n_i},
\]

where \(n_i\) is the number of cows in each trial and \(k\) is the number of studies being assimilated (Hedges and Olkin, 1985).

The variance of the effect size for each trial is calculated (Hedges and Olkin, 1985):

\[
Var^2(ES) = \frac{n_e + n_c}{n_e n_c} + \frac{ES^2}{2(n_e + n_c)},
\]

and 95% confidence intervals for the ES of each trial and for \(ES_o\):

\[
95\% \ CI = ES \pm 1.96\sqrt{Var^2(ES)}.
\]

Weights are calculated

\[
W_i = \frac{1}{\text{var}_i},
\]

where \(W_i\) is the weight assigned to each study, and \(\text{var}_i\) is the variance.

The methods used in meta-analysis of continuous data, particularly in reference to nutrition studies in animals, have been explored by St-Pierre (2001) and Sauvant et al. (2008). Sauvant et al. (2008) explore the use of mixed models and regression analysis in meta-analysis, and the methods used in these papers should be examined along with those used by Duffield et al. (2008a) for continuous data. The use of analytic methods that allow an assessment of heterogeneity is a substantial consideration. Duffield et al. (2008b) reported significant effects of monensin on decreasing DMI and milk fat percentage in lactating dairy cows. The analyses also illustrated the extremely consistent effect of monensin on DMI, yet a very variable effect on milk fat percentage, which led to further exploration of sources of that heterogeneity, including dietary factors. Both St-Pierre (2001) and Sauvant et al. (2008) examine the methods and effects of weighting studies. An important consideration for meta-analysis of nutrition studies is the use of crossover or Latin square designs. These studies pose particular challenges in animal health or production. Concerns with these studies include the potential for carry-over of effects from one period to another, suitability of the condition to study concerning follow-up periods or stage of lactation, a
potential for exclusion bias, and effects of stage of lactation. The carry-over effect in a Latin square design is illustrated in a study of feeding fish oil and monensin to lactating dairy cattle (Cant et al., 1997). In this investigation, a 3-wk period for washout/adaptation was used. However, there was clearly carry-over from the fish oil treatment to the monensin group because 2 of 4 cows fed the monensin only treatment had detectable docosahexaenoic acid in the milk fatty acid profile. After checking the treatment order, we confirmed that the 2 cows had previously been fed fish oil before receiving the monensin only treatment. We [C. J. Sniffen (Fencrest LLC, Holderness, NH), M. B. de Ondarza (Paradox Nutrition LLC, West Chazy, NY), and I. J. Lean; unpublished data] recently identified very substantial differences in weighting of Latin square and randomized controlled trials examining the effects of diet on milk protein production of dairy cattle. This difference reflected the much smaller variance in milk protein production of cows in Latin square studies. A more conservative weighting based on the square root of the number of animals in studies may be the most appropriate way to weight studies that include Latin square designs. However, serious consideration should be given to the appropriateness of including Latin square designs in pooled analyses.

HETEROGENEITY

Arguably, the greatest benefit of conducting meta-analysis is to examine sources of heterogeneity, if present, among studies. If heterogeneity is present, the summary measure must be interpreted with caution (Greenland, 1987). When heterogeneity is present, one should question whether and how to generalize the results. Understanding sources of heterogeneity will lead to more effective targeting of prevention and treatment strategies and will result in new research topics being identified. Part of the strategy in conducting a meta-analysis is to identify factors that may be significant determinants of subpopulation analysis or covariates that may be appropriate to explore in all studies. To understand the nature of variability in studies, it is important to distinguish between different sources of heterogeneity. Variability in the participants, interventions, and outcomes studied has been described as clinical diversity (Cochrane Collaboration, 2008), and variability in study design and risk of bias has been described as methodological diversity (Cochrane Collaboration, 2008). Variability in the intervention effects being evaluated among the different studies is known as statistical heterogeneity and is a consequence of clinical or methodological diversity, or both, among the studies. Statistical heterogeneity manifests itself in the observed intervention effects varying by more than the differences expected among studies that would be attributable to random error alone. Usually, in the literature, statistical heterogeneity is simply referred to as heterogeneity.

Clinical variation will cause heterogeneity if the intervention effect is modified by the factors that vary across studies; most obviously, the specific interventions or participant (animal) characteristics that are often reflected in different levels of risk in the control group when the outcome is dichotomous. In other words, the true intervention effect will differ for different studies. Differences between studies in terms of methods used, such as use of blinding or differences between studies in the definition or measurement of outcomes, may lead to differences in observed effects. Significant statistical heterogeneity arising from differences in methods used or differences in outcome assessments suggests that the studies are not all estimating the same effect, but does not necessarily suggest that the true intervention effect varies. In particular, heterogeneity associated solely with methodological diversity indicates that studies suffer from different degrees of bias. Empirical evidence suggests that some aspects of design can affect the result of clinical trials, although this may not always be the case.

The scope of a meta-analysis will largely determine the extent to which studies included in a review are diverse. Meta-analysis should be conducted when a group of studies is sufficiently homogeneous in terms of animals involved, interventions, and outcomes to provide a meaningful summary. However, it is often appropriate to take a broader perspective in a meta-analysis than in a single clinical trial. Combining studies that differ substantially in design and other factors can yield a meaningless summary result, but the evaluation of reasons for the heterogeneity among studies can be very insightful. It may be argued that these studies are of intrinsic interest on their own, even though it is not appropriate to produce a single summary estimate of effect.

Statistical Assessment of Heterogeneity

Variation among the trial level ES or risk ratios are usually assessed using Cochrane’s $Q$ statistic, a chi-squared ($\chi^2$) test of heterogeneity. The null hypothesis is that the effect of treatment was the same across $k$ trials, and the null hypothesis is rejected if the heterogeneity test statistic was greater than a critical value that separated the upper 10% of a $\chi^2$ distribution with $(k - 1)$ df. This test has relatively poor power to detect heterogeneity among small numbers of trials (Egger
and Smith, 2003); consequently, an α level of 0.10 is used to test hypotheses.

Heterogeneity of results among trials is better quantified using the $I^2$ statistic (Higgins et al., 2003), which describes the percentage of total variation across studies that is due to heterogeneity rather than chance. Where $Q$ is the $\chi^2$ heterogeneity statistic and $k$ is the number of trials, $I^2$ is calculated as

$$I^2(\%) = \frac{Q - (k - 1)}{Q} \times 100.$$  \[11\]

Uncertainty intervals for $I^2$ (dependent on $Q$ and $k$) are calculated using the method described by Higgins and Thompson (2002). Negative values of $I^2$ are put equal to zero, consequently $I^2$ lies between 0 and 100%. A value $>50\%$ may be considered substantial heterogeneity (Higgins and Thompson, 2002). This statistic is less influenced by the number of trials compared with other methods used to estimate the heterogeneity and provides a logical and readily interpretable metric.

**Approaches to Heterogeneity**

Given that there are several potential sources of heterogeneity in the data, several steps should be considered in the investigation of the causes. Although random-effects models, either frequentist or Bayesian, are appropriate, we consider that it is still very desirable to examine the data to identify sources of heterogeneity and to take steps to produce models that have a lower level of heterogeneity, if appropriate. Further, if the studies examined are highly heterogeneous, we consider that it is not appropriate to present an overall summary estimate, even when random effects models are used. As Petiti (1994) notes, statistical analysis alone will not make contradictory studies agree; critically, however, one should use common sense in decision-making. Despite heterogeneity in responses, if all studies had a positive point direction and the pooled confidence interval did not include zero, it would not be logical to conclude that there was not a positive effect, provided that sufficient studies and subject numbers were present. The appropriateness of the point estimate of the effect is much more in question.

L’Abbé plots (L’Abbé et al., 1987) are used to examine responses in dichotomous outcomes. The L’Abbé plots are used to show the variation in observed results by plotting the event probability in the treatment group on the vertical axis and that in the control group on the horizontal axis. The visual presentation is easy to assimilate, because the source of variability in response can be from deviation from the expected results, as represented by the null result straight line, by the control group, the treated group, or both. The information provided by these deviations can be used to identify sources of heterogeneity. For instance, if a L’Abbé plot reveals that there is a great variability in outcome from control groups but a similar outcome in intervention groups, heterogeneity in the treatment effect across individual studies may be at least partially explained by differences in the control populations. In this case, it may be more fruitful to investigate differences in control groups.

Galbraith plots (Galbraith, 1988) provide a graphical display to obtain a visual impression of the amount of heterogeneity from a meta-analysis. For each trial, the $z$ statistic is plotted against the reciprocal standard error. The unweighted regression line constrained through the origin, with its 95% confidence interval, has a slope equal to the overall log risk ratio, or log odds ratio, or log hazard ratio in a fixed effects meta-analysis. The position of each trial on the horizontal axis gives an indication of the weight allocated to it in a meta-analysis. The position on the vertical axis gives the contribution of each trial to the $Q$ statistic for heterogeneity. In the absence of heterogeneity, we could expect all the points to lie within the confidence bounds (positioned 2 units over and below the regression line). Figure 4 from Rabiee et al. (2005) shows the heterogeneity in Ovsynch and prostaglandin F2α data using a Galbraith plot ($I^2 = 73.3\%, P < 0.0001$). Clearly, many of these studies were heterogeneous. Identifying characteristics of these studies is critical to the investigation of sources of heterogeneity.

Several approaches have been used to examine the data for the causes of heterogeneity. Meta-regression is a weighted regression of the study results based on factors of interest that may influence the responses of subjects to treatment (Dohoo et al., 2003a,b). Sanchez et al. (2004) used meta-regression to examine differences in statistical approaches and source of publication of studies examining the effects of anthelmintic treatments on milk production as sources of heterogeneity. These factors and the presence or absence of formal randomization and blinding were evaluated as indications of study quality. Rabiee et al. (2008) conducted a detailed evaluation of study quality on responses to yeast supplementation and found a high level of consistency in ES between published and unpublished studies. Other factors examined by Sanchez et al. (2004) included methodological differences among studies; for example, anthelmintic versus endectocides, timing of treatment in regard to stage of lactation, whole or partial herd treatments, and parity of cattle treated. Duffield et al. (2008b) were able to explore a large number of factors that influenced responses to
monensin treatment including some that had not previously been evaluated as sources of variation in response, specifically consistency of delivery system, presence of ruminal peptides on milk protein yield, and impact of dietary fatty acid precursors on milk fat percentage in a multivariate study. These findings and others, such as those of Oetzel (1991), reinforce a key strength of meta-analysis, which is the generation of new hypotheses and the capacity to test these.

Stratified analyses can also be used to reduce heterogeneity. Morgan and Lean (1993) used this approach to reduce, but not eliminate, heterogeneity in the GnRH treatment at insemination data. In this case, the dose of GnRH and insemination number influenced treatment responses. Burton and Lean (1995) found that the weighted average reduction in days open between treated and control cows was 2.6 d for trials with abnormal cows and 3.3 d for trials including normal and abnormal cows for studies examining the use injections of prostaglandin F₂₀ after calving.

Sensitivity analyses have also been used to examine the effects of studies identified as being aberrant concerning conduct or result, or being highly influential in the analysis. Rabiee et al. (2004, 2005) used this method in exploring sources of heterogeneity in the controlled internal drug-releasing insert (CIDR) and Ovsynch data. Baker and Jackson (2008) have proposed a method that reduces the weight of studies that are outliers in meta-analyses. All of these methods for examining heterogeneity have merit, and the variety of methods available reflects the importance of this activity.

**PUBLICATION BIAS**

Although the intent of a meta-analysis is to find and assess all studies meeting the inclusion criteria, it is not always possible to obtain these. A critical concern is the papers that may have been missed. There is good reason to be concerned about this potential loss because studies with significant, positive results are more likely to be published and, in the case of interventions with a commercial value, to be promoted, than studies with nonsignificant or “negative” results. Studies that produce a positive result, especially large studies, are more likely to have been published and, conversely, there has been a reluctance to publish small studies that have nonsignificant results. Further, publication bias is not solely the responsibility of editorial policy as there is reluctance among researchers to publish results that were either uninteresting or are not randomized (Dickersin and Berlin, 1992). There are, however, problems with simply including all studies that have failed to meet peer-review standards. All methods of retrospectively dealing with bias in studies are imperfect.
Managing Publication Bias

It is important to examine the results of each meta-analysis for evidence of publication bias. An estimation of likely size of the publication bias in the review and an approach to dealing with the bias are inherent to the conduct of many meta-analyses. Several methods have been developed to provide an assessment of publication bias; the most commonly used is the funnel plot. The funnel plot was developed by Light and Pillemer (1984) and provides a graphical evaluation of the potential for bias. If publication bias is not present, the funnel plot is expected to be symmetrical, as shown in Figure 5. In this example from Duffield et al. (2008a), there is a symmetrical pattern of distribution around the point estimate, indicated by the point of the funnel on the y-axis. In a study in which there is little publication bias, larger studies (as indicated by the size of the circles in the plot) tend to cluster closely to the point estimate. As studies become less precise (i.e., have a higher standard error), the results of the studies can be expected to be more variable and are scattered to both sides of the more precise larger studies. Figure 5 shows that the smaller, less precise studies are, indeed, scattered to both sides of the point estimate of effect and that these seem to be symmetrical showing little evidence of publication bias.

In contrast to Figure 5, Figure 6 from Duffield et al. (2008a) shows evidence of publication bias. There is evidence of the possibility that studies using smaller numbers of subjects and showing an increase in 3-hydroxybutyrate concentrations in blood were not published. Early evidence that monensin reduced concentrations of 3-hydroxybutyrate was strong and those conducting small studies who found an increase in 3-hydroxybutyrate may have been less likely to publish.

Asymmetry of funnel plots is not solely attributable to publication bias, but may also result from clinical heterogeneity among studies. Sources of clinical heterogeneity include differences in control or exposure of subjects to confounders or effect modifiers, or methodological heterogeneity between studies; for example, a failure to conceal treatment allocation. There are several statistical tests for detecting funnel plot asymmetry; for example, Begg’s rank correlation test and Egger’s linear regression test, but these do not have considerable power and are rarely used.

The adjusted rank correlation (Begg’s method) is used to assess the correlation between estimates and their variances. The deviation of Spearman’s $\rho$ values from zero provides an estimate of funnel plot asymmetry. Positive values indicate a trend toward higher effect estimates in studies with smaller sample size (Begg and Mazumdar, 1994). The Egger’s test is used to determine whether there is an association between study results and their precision. A regression analysis is conducted using the following equation:

$$SND = a + b \times \text{SE}(d)^{-1},$$

where SND is the standard normal deviate defined as diagnostic $d$ divided by its standard error $\text{SE}(d)$, $a$ is the intercept, and $b$ is the slope. The intercept value ($a$) provides an estimate of asymmetry of funnel plot, with positive values ($a > 0$) indicating a trend toward higher levels of test accuracy in studies with smaller sample size (Egger et al., 1997).

Contour-Enhanced Funnel Plots. Contour-enhanced funnel plots have been proposed by Peters et al. (2008) to include contour lines corresponding...
to statistical significance \((P = 0.01, 0.05, 0.1)\). This approach allows the statistical significance of study estimates and areas in which studies are perceived to be missing to be considered. The contour-enhanced funnel plots may help to differentiate asymmetry caused by publication bias from that due to other factors. For example, if studies appear to be missing in areas of statistical nonsignificance, then this adds credence to the possibility that the asymmetry is caused by publication bias. Conversely, if the supposed missing studies are in areas of higher statistical significance, this suggests that the observed asymmetry may be more likely to be due to factors other than publication bias; for example, variable study quality or a failure to publish findings that were not statistically significant (Figure 7). In Figure 7, the 2 studies that show very positive effects that lie to the right of the 0.01 line represent strong positive responses that are unlikely to be matched by negative studies. If there are no statistically significant studies, then publication bias may not be a plausible explanation for funnel plot asymmetry (Ioannidis and Trikalinos, 2007).

**Fail-Safe N.** Rosenthal (1979) suggested a means of assessing the potential for publication bias to have influenced the results of a meta-analysis by calculating a “fail-safe N.” This is the number of additional “negative” studies, specifically studies in which the intervention effect was zero, that would be required to increase the \(P\)-value for the meta-analysis above 0.05 (Rosenthal, 1979). The estimate of fail-safe N depends greatly on the mean intervention effect that is assumed for the unpublished studies (Iyengar and Greenhouse, 1988), and available methods lead to widely varying estimates of the number of additional studies (Becker, 2005). The method also runs against the strong principle of systematic reviews, that one should concentrate on the size of the estimated intervention effect and associated confidence intervals, rather than on whether the \(P\)-value reaches a particular, arbitrary threshold. Similar methods for ES estimates have also been proposed (Orwin, 1983). Given the desirability of maintaining focus on confidence interval estimation, this and related methods are not recommended for the detection of publication bias.

**Trim and Fill.** A rank-based data augmentation technique, the “trim and fill” method can be used to estimate the number of missing studies and to produce an adjusted estimate of test accuracy by imputing suspected missing studies (Duval and Tweedie, 2000). Smaller studies are omitted until the funnel plot is symmetrical (trimming). The trimmed funnel plot is used to estimate the true center of the funnel, and then the omitted studies and their missing counterparts around the center are replaced (filling). This provides an estimate of the number of missing studies and adjusted treatment effect, including the filled studies. Sanchez et al. (2004) used a trim and fill method to explore the potential for missing data to influence the effects of anthelmintic treatments on milk production.

**INFLUENTIAL STUDY ANALYSIS**

Influential studies are of interest in statistical analysis. Large, well-conducted studies are important to identify, as are studies with a profoundly influential effect on the point estimate. In the case of a large study that has been well conducted but differs in estimate of effect to smaller studies, the key concern is whether the large study provides a stronger evidential base than the other studies conducted. In the case of the small study with strong effect, it is often illuminating to examine the study for differences in study design or analysis that may have influenced the outcome (Tobias, 1999). The following influential study analyses from Duffield et al. (2008a,c) show a series of studies in which very few studies were influential and one in which there is clear evidence of a study with a strong effect (Figures 8 and 9).

**Replicability of Meta-Analyses**

Chalmers et al. (1987) examined the replicability of meta-analyses conducted on 18 topics in medical re-
search. For the 52 separate meta-analyses studied, 3 pairs of studies differed in direction; that is, one study in the pair favored treatment or control and the other found no effect. All but 7 of the remainder (n = 42) agreed on the direction and significance of effect. The 7 remaining studies agreed on direction, but not significance.

The studies of Gross et al. (1999) and Sanchez et al. (2004), who studied the effects of anthelmintic treatments on milk production, agreed in both significance and direction. Gross et al. (1999) estimated a 0.63 kg of milk per cow per day response to anthelmintic treatment, whereas Sanchez et al. (2004) found an adjusted estimate of 0.35 kg of milk per cow per day response to treatment.

Dietary and other risk factors for milk fever have been examined in 4 meta-analyses providing multivariable models and one model providing a univariable estimate (Charbonneau et al., 2006). Two of the studies used identical data, and Oetzel (1991), Enevoldsen (1993), and Lean et al. (2006) developed 2 models of similar statistical fit using an expanded milk fever data set including the original data from Oetzel (1991). Oetzel (1991) found a significant univariable relationship between DCAD and risk of milk fever, as did Charbonneau et al. (2006), albeit with a slightly different definition for DCAD. The multivariable model of Lean et al. (2006) supported that finding. The consistency of estimate of effect for variables retained within the multivariable models in Table 2 is reasonably high, despite a finding that both magnesium and exposure acted to confound the coefficients for calcium in the models developed by Lean et al. (2006). Estimates of increased risk of milk fever with CP and lactation number from the Oetzel (1991) data are consistent with unpublished models that included these factors investigated by Lean et al. (2006), but are not consistent with the effect estimated by Enevoldsen (1993). Despite the latter finding,
CONCLUSIONS

The traditional basis of medical practice has been changed by the use of randomized, blinded, multicenter clinical trials and meta-analysis, leading to the widely used term “evidence-based medicine.” Leaders in initiating this change have been the Cochrane Collaboration who have produced guidelines for conducting systematic reviews and meta-analyses (Cochrane Collaboration, 2008) and texts demonstrating the methods used (Egger et al., 2001) and results found (Sinclair and Bracken, 1992). Animal and veterinary sciences have, unfortunately, been slower to adopt these approaches despite excellent base materials for these analyses. We believe there will be considerable growth in this area of research in the future.

Results of studies to date have demonstrated the merit of this approach. Traditional review methods recommended that calcium concentrations be maintained at 1.0 to 1.2% of dietary intake before calving (Oetzel, 2000). These concentrations appear to increase milk fever risk, whereas meta-analyses (Oetzel, 1991; Lean et al., 2006) found a quadratic risk with low (<0.6% of DMI) and possibly high concentrations (>1.5% of DMI) reducing risk of milk fever. The use of GnRH at insemination was considered ineffective (Wright and Malmo, 1992), whereas meta-analysis found significant positive effects (Morgan and Lean, 1993). The effect of treatment of cattle with anthelmintics on milk production was considered in need of clarification (Vercruysse and Claerebout, 2001); the very consistent meta-analyses of Gross et al. (1999) and Sanchez et al. (2004) provide evidence of a positive response to these treatments and factors that influence the magnitude of the response. The meta-analysis of the health effects of rbST in dairy cattle was influential in the Canadian decision to deny registration of the product Dohoo et al. (2003b). Duffield et al. (2008a,b,c) identified several new understandings regarding use of monensin to modify rumen function.

No single study, whether meta-analytic or not, will provide the definitive understanding of responses to treatment, diagnostic tests, or risk factors influencing disease. Despite this limitation, meta-analytic approaches have demonstrable benefits in addressing the limitations of study size, can include diverse
Table 2. A comparison of models developed that examine the effect of macromineral concentrations in the precalving diet and other factors on the risk of milk fever1,2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Logistic regression coefficient</th>
<th>Random effects logistic regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oetzel (1991)</td>
<td>Lean et al. (2006); model 1</td>
</tr>
<tr>
<td></td>
<td>Enevoldsen (1993)</td>
<td>Lean et al. (2006); model 2</td>
</tr>
<tr>
<td>Ca, %</td>
<td>6.79</td>
<td>3.58</td>
</tr>
<tr>
<td>Ca, %2</td>
<td>−2.92</td>
<td>−2.03</td>
</tr>
<tr>
<td>DCAD</td>
<td>NA</td>
<td>0.02</td>
</tr>
<tr>
<td>Mg, %</td>
<td>NA</td>
<td>−5.05</td>
</tr>
<tr>
<td>S, %</td>
<td>−13.06</td>
<td>NA</td>
</tr>
<tr>
<td>K, %</td>
<td>NS</td>
<td>NA</td>
</tr>
<tr>
<td>Na, %</td>
<td>1.25</td>
<td>1.85</td>
</tr>
<tr>
<td>Cl, %</td>
<td>NS</td>
<td>3.78</td>
</tr>
<tr>
<td>CP, %</td>
<td>0.14</td>
<td>1.85</td>
</tr>
<tr>
<td>S, %2</td>
<td>NP</td>
<td>13.22</td>
</tr>
<tr>
<td>Cl, %2</td>
<td>NP</td>
<td>−6.05</td>
</tr>
<tr>
<td>K, %2</td>
<td>NP</td>
<td>4.07</td>
</tr>
<tr>
<td>Lactation number</td>
<td>0.36</td>
<td>NA</td>
</tr>
<tr>
<td>Exposure to transition diet (d)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

1NP = not provided; NA = not applicable; NS = not significant; Sig = significant, but a different basis for assessment.

2 Another factor included in all models was breed of cattle. The Enevoldsen (1993) model included also terms for DM content of the diet and the quadratic effect of this, ether extract (fat %) and the quadratic effect of this, energy density of the diet, crude fiber percentage and the quadratic effect of this, interactions between DM content and the chloride percentage and DM and the calcium percentage.

populations, provide the opportunity to evaluate new hypotheses, and are more valuable than any single study contributing to the analysis. The conduct of the studies is critical to the value of a meta-analysis and the methods used need to be as rigorous as any other study conducted. Methods and insights provided in this paper should assist those seeking to evaluate or conduct meta-analyses.

REFERENCES


