STATISTICS IN MEDICINE
Statist. Med. **18**, 321–359 (1999)

TUTORIAL IN BIOSTATISTICS
META-ANALYSIS: FORMULATING, EVALUATING, COMBINING, AND REPORTING

SHARON-LISE T. NORMAND *

Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A., and Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

SUMMARY

Meta-analysis involves combining summary information from related but independent studies. The objectives of a meta-analysis include increasing power to detect an overall treatment effect, estimation of the degree of benefit associated with a particular study treatment, assessment of the amount of variability between studies, or identification of study characteristics associated with particularly effective treatments. This article presents a tutorial on meta-analysis intended for anyone with a mathematical statistics background. Search strategies and review methods of the literature are discussed. Emphasis is focused on analytic methods for estimation of the parameters of interest. Three modes of inference are discussed: maximum likelihood; restricted maximum likelihood, and Bayesian. Finally, software for performing inference using restricted maximum likelihood and fully Bayesian methods are demonstrated. Methods are illustrated using two examples: an evaluation of mortality from prophylactic use of lidocaine after a heart attack, and a comparison of length of hospital stay for stroke patients under two different management protocols. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies. The objectives of a meta-analysis can be several-fold. By combining information over different studies, an integrated analysis will have more statistical power to detect a treatment effect than an analysis based on only one study. For example, Hine *et al.*[1] conducted a meta-analysis of death rates in randomized controlled trials in which prophylactic lidocaine was administered to patients with proved or suspected acute myocardial infarction. Table I describes mortality at the end of the assigned treatment period for control and intravenous lidocaine treatment groups for six studies. The unadjusted total mortality rates were 6·6 per cent (37/557) in the lidocaine group and 3·8 per cent (21/549) in the control group. The question of interest is whether there is a detrimental effect of lidocaine. Because the studies were conducted to compare rates of arrhythmias following a heart attack, the studies, taken individually, are too small to detect important differences in mortality rates.

* Correspondence to: Sharon-Lise T. Normand, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115, U.S.A.

Contract/grant sponsor: National Cancer Institute
Contract/grant number: CA-61141

CCC 0277–6715/99/030321–39 $17.50

Received January 1998
Accepted April 1998

Table I. Prophylactic use of lidocaine after a heart attack: evaluating mortality from prophylactic use of lidocaine in acute myocardial infarction. Source: reference 1

Source	Number randomized		Number dead	
	Lidocaine	Control	Lidocaine	Control
1. Chopra *et al.*	39	43	2	1
2. Mogensen	44	44	4	4
3. Pitt *et al.*	107	110	6	4
4. Darby *et al.*	103	100	7	5
5. Bennett *et al*	110	106	7	3
6. O'Brian *et al.*	154	146	11	4
Total	557	549	37	21

Figure 1 displays the individual study mortality risk differences and corresponding 95 per cent confidence intervals. Note that the first five studies indicate no effect of lidocaine on mortality; the sixth and largest study, although not indicating a statistically significant effect, does provide some evidence of a detrimental effect of lidocaine.

When several studies have conflicting conclusions, a meta-analysis can be used to estimate an *average* effect or to identify a subset of studies associated with a beneficial effect. For example, the Cochran Database of Systematic Reviews[2] collected information on all trials where specialist inpatient stroke care was compared to the conventional non-specialist care. Table II lists nine studies, and for each, the average length of stay (LOS) during the acute hospital admission and its standard deviation. The central hypothesis of interest is whether specialist stroke unit care will result in a shorter length of hospitalization compared to routine management. Figure 2 displays the difference in the average length of stay for stroke patients managed in specialty units from the average length of stay for the group managed in the routine manner. Four of the nine studies (studies 1, 3, 4 and 8) resulted in statistically significant shorter stays in the specialty units compared to those in the general wards.

A systematic approach to synthesizing information can provide estimates of the degree of benefit from a particular therapy and whether the benefit depends upon specific characteristics of the studies. This latter question capitalizes on the differences *across* studies; it is possible to test whether there are differences in the size or direction of the treatment effect associated with study-specific variables. For example, length of stay may be substantially shorter in publicly-owned hospitals for specialty managed stroke patients than for routinely managed patients but no different in for-profit hospitals. Moreover, it may be of more interest to find a particularly effective treatment than in determining whether all studies, on average, involve effective treatments.

This tutorial is intended for readers with a mathematical statistics background. Meta-analysis in which the individual study summary statistics are Normal variables arising from two-arm studies (such as those introduced above) are considered. For moderately large study sizes, the summary statistics should be asymptotically normally distributed. The situation in which there are multiple dependent variables measured on each subject requires an additional level of data synthesis. For example, if each study involves measuring the results of several tests per subject, then in order to make efficient use of within-subject information, multivariate methods should be utilized. Typically,

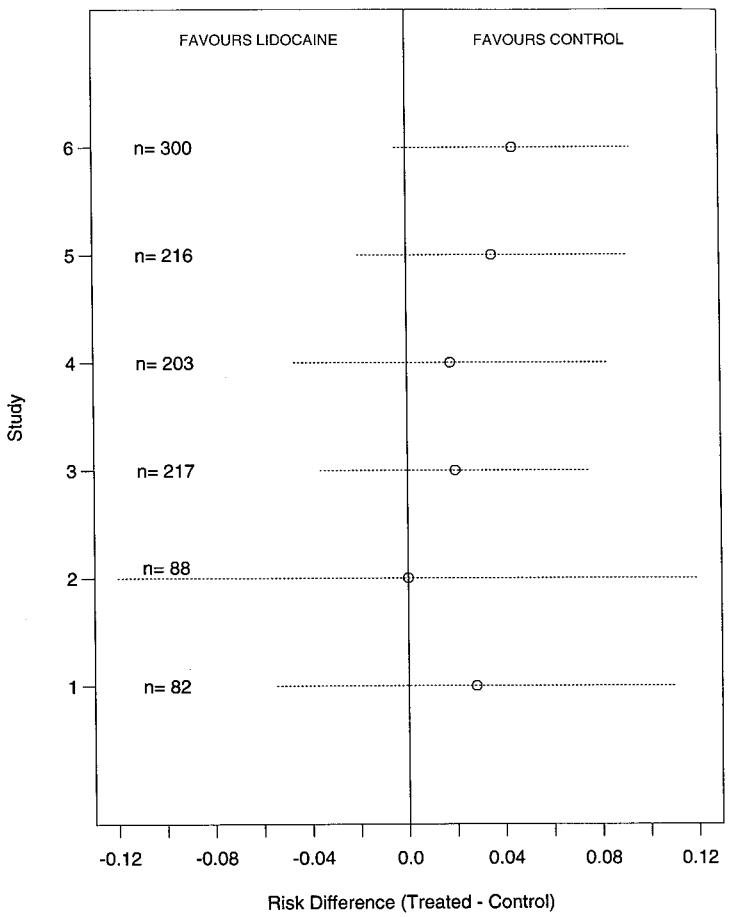

Figure 1. Prophylactic lidocaine after a heart attack. The x-axis displays the risk difference, $d_i = \hat{p}_{Ti} - \hat{p}_{Ci}$, and corresponding 95 per cent confidence intervals $\left(s^2_{d_i} = \frac{\hat{p}_{Ti}\hat{q}_{Ti}}{n_{Ti}} + \frac{\hat{p}_{Ci}\hat{q}_{Ci}}{n_{Ci}}\right)$; the y-axis indicates the study and total sample size

however, this is not an easy task and space limitations do not permit an adequate discussion of methods to combine multiple summary statistics across several studies.

In Section 2 issues relating to the definition of the study objectives, the domain of the literature search, and the search strategies are reviewed. Section 3 briefly describes how to evaluate the literature once it is retrieved. Section 4 summarizes several analytic methods for combining information across studies including the choice of outcome and estimation in fixed-effects and random-effects models. Because software accommodating estimation in a random-effects model is not streamlined, two packages for conducting inference in a random-effects model are described in Section 5. Section 6 identifies features of the meta-analysis that should be reported. Finally, methods are illustrated in Section 7 using the lidocaine data and in Section 8 using the stroke data. Throughout this tutorial, studies comprising the meta-analysis are denoted *primary* studies.

Table II. Specialist care for stroke patients from nine studies: comparing specialist multidisciplinary team care for managing stroke inpatients with routine management in general medical wards. Source: reference 2

Source	Specialist care			Routine management		
	N	Mean LOS	SD	*N*	Mean LOS	SD
1. Edinburgh	155	55·0	47·0	156	75·0	64·0
2. Orpington-Mild	31	27·0	7·0	32	29·0	4·0
3. Orpington-Moderate	75	64·0	17·0	71	119·0	29·0
4. Orpington-Severe	18	66·0	20·0	18	137·0	48·0
5. Montreal-Home	8	14·0	8·0	13	18·0	11·0
6. Montreal-Transfer	57	19·0	7·0	52	18·0	4·0
7. Newcastle 1993	34	52·0	45·0	33	41·0	34·0
8. Umea 1985	110	21·0	16·0	183	31·0	27·0
9. Uppsala 1982	60	30·0	27·0	52	23·0	20·0
Total	548			610		

LOS = length of stay measured in days; SD = standard deviation.

Text references on meta-analysis include Cooper and Hedges,[3] Hedges and Olkin[4] and Rosenthal.[5] Cook *et al.*[6] describe guidelines for reviewing randomized controlled trials. An analogy is worth bearing in mind throughout: conducting a meta-analysis is conceptually *no different* than conducting primary research. It is multidisciplinary and therefore requires a research team comprising several experts: a subject-matter specialist (for example, neurologist, cardiologist); a biostatistician to help in the design and analytic aspects of the research; an information librarian to provide guidance regarding printed sources and databases; data coders to translate the data from the literature into the data base, and a group of subject-matter specialists to aid in judging the relevance of the retrieved documents.

2. FORMULATING

2.1. Beginning a meta-analysis

As in primary research, a meta-analysis begins with a well-formulated question and design. What are the *study objectives*? Is the objective of the study to validate results in a broader population? For example, is lidocaine prophylaxis during a heart attack related to any mortality effect? Or is the goal to guide *new* studies? For example, which aspects of stroke care are beneficial?

What are the *operational definitions* of the research outcome (treatment period mortality or total mortality), the treatment or intervention (bolus $\geqslant$ 50 mg of lidocaine followed by continuous infusion greater than 1·0 mg/min for at least 24 hours or single-dose therapy), and the population (patients who have a *confirmed* heart attack or patients with a *suspected* heart attack).

What types of *designs* will be included in the search? Will only randomized trials testing the research hypothesis be included or will the results from non-experimental studies be permitted? Will randomized trials with poor compliance be included?

The answers to these questions impact on the methods of review, the modes of statistical inference, and the interpretation of the results. If interest is centred on making inferences for the very populations that have been sampled, then the treatment levels are considered *fixed* and the

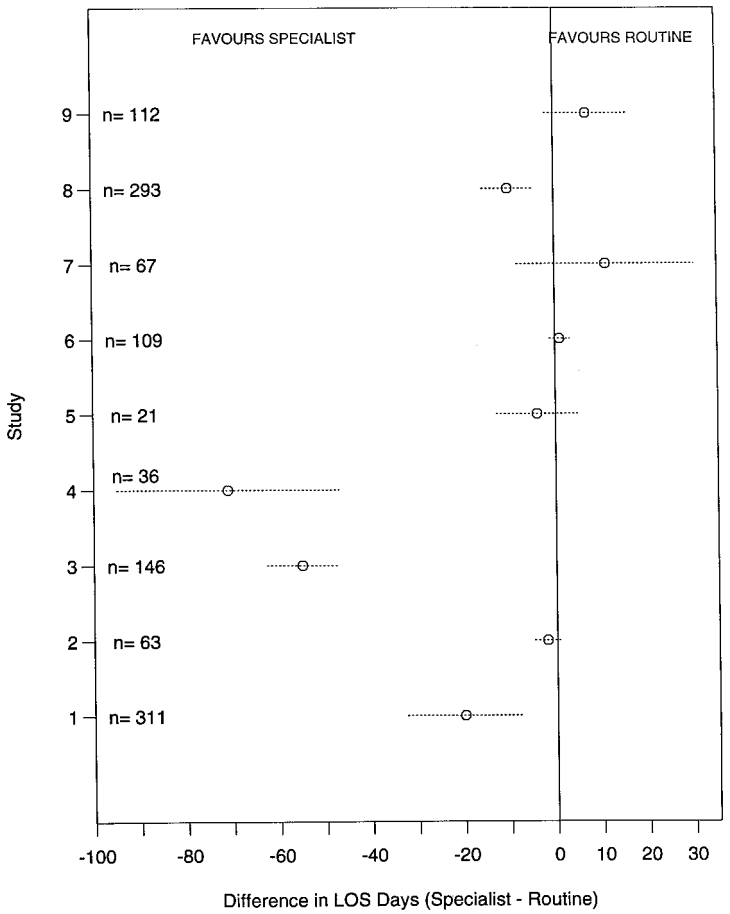

Figure 2. Stroke example. x-axis displays $\bar{x}_{Ti} - \bar{x}_{Ci}$, the difference in average length of stay, measured in days, for each of the nine studies and their corresponding 95 per cent confidence intervals $\left(s_i^2 = s_{pi}^2 \left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}\right)\right.$ where $\left. s_{pi}^2 = \left\{\frac{(n_{Ti}-1)s_{Ti}^2 + (n_{Ci}-1)s_{Ci}^2}{n_{Ti}+n_{Ci}-2}\right\}\right)$. Shorter lengths of stay are assumed to reflect better care management

only source of uncertainty is that resulting from the sampling of people into studies. This type of variation may be characterized as within-study variation that is a function of the number of patients in the primary study and the variability in the patient responses within the primary studies. In this case a *fixed-effects* (Section 4.2.1) model would be used for statistical inferences. Inferences are similar to those made when performing an analysis of variance (ANOVA) when there is no inter-study variation in the mean outcome. The intuition underpinning the fixed-effects model is that other levels of treatments are sufficiently like those in the sample of primary studies that inferences would be the same. The population to which generalizations are to be made consists of a set of studies having identical characteristics and study effects. Thus, in the stroke management example, inferences regarding the reduction in bed days in a hospital based on a fixed-effects model are applicable to management teams identical to those in the nine primary studies.

On the other hand, if inferences are to be generalized to a population in which the studies are permitted to have different effects and different characteristics, then a *random-effects* (Section 4.2.2) model would be appropriate. The intuition underpinning random-effects models is that because there are many different approaches to conducting a study by perturbing the design in a small way, then there are many different potential treatment effects that could arise. This situation corresponds to an ANOVA model in which there is inter-study variation in the mean outcome in addition to the within-study variation. Thus, the population in a random-effects model is the one in which there are infinitely many possible populations. Inferences regarding the effect of lidocaine based on a random-effects model applies to the population that would be formed if additional studies were sampled in a manner similar to that used to obtain the six primary studies.

There are conceptual difficulties linked to both the fixed-effects and random-effects points of view. In both models, it may be difficult to characterize precisely the universe to which we are inferring. In the random-effects model, the universe may too big to imagine and in the fixed-effects model, too small to be of any practical importance. There is a long debate as to the choice of appropriate model that cannot be adequately covered in this tutorial. What should be noted is that it is almost *always* reasonable to believe that there is some between-study variation and few reasons to believe it is zero. However, if all the lidocaine studies indicate that lidocaine is related to increased mortality then it may be of little concern that some studies favour it more strongly than others. On the other hand, when studies conflict, such as in the stroke example, it is difficult to ignore the between-study variation.

2.2. The domain of the literature search

Once the researcher has established the goals of the meta-analysis, an ambitious literature review needs to be undertaken, the literature obtained, and then summarized. Sources to be searched include the published literature, unpublished literature, uncompleted research reports, and work in progress. The meta-analyst begins with searches of regular bibliographic reports: citation indexes (for example, the Social Sciences Citation Index) and abstract databases (for example, Mental Health Abstracts) provide information regarding published reports. These publications are retrieved and the references therein are searched for more references, the new publications retrieved, and the process is repeated again and again.

Reliance on only published reports lead to *publication bias* – the bias resulting from the tendency to selectively publish results that are statistically significant. Study design features, such as small sample size or failure to randomize, may be positively associated with the bias. As a first step towards eliminating publication bias, the meta-analyst needs to obtain information from unpublished research. For example, *ERIC* (Educational Resources Information Center) is a database that includes references to unpublished reports and conference papers in addition to published works. The *NTIS* (National Technical Information Service) is a bibliographic database summarizing completed research sponsored by more than 600 U.S. federal agencies; *UNIVRES* (Directory of Federally Supported Research in Universities) lists approximately 180,000 university-based research projects sponsored in Canada.

Clinical research and clinical trials registers are another valued source of information. These data sources contain information on all initiated studies and are typically maintained by their funding institution or by groups of individuals with a particular interest in the subject area. For example, the Neurosurgery Clinical Trials Registry lists information on all completed, active and planned clinical trials in neurosurgery; the International Registry of Vision Trials is a univer-

sity funded database containing information of completed and active opthamology and optometry trials.

Unpublished dissertations and master's theses can also be searched in databases; for example, the database produced by the University Microfilms International Dissertation Service (Ann Arbor, MI) contains citations for theses published as early as 1961. Early dissemination of scientific results often occurs at conferences and so the meta-analyst must also obtain this literature (for example, conference indexes such as British Library Lending Division Conference Index).

In summary, the literature search needs to be a well-formulated and co-ordinated effort involving several researchers. It is well-advised to seek the guidance of an information scientist to oversee this aspect of the meta-analysis.

2.3. Quantitative aspects of the search

Although the meta-analyst wants to perform as complete a search as possible, it is clearly not feasible to obtain *every* piece of literature that is related to the research topic. Two concepts in information retrieval can be used to describe the success of the search process: *recall* and *precision*. Recall is defined as

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number that should be retrieved}} \times 100$$

and measures the success of the retrieval process with a higher per cent recall corresponding to a more successful search. Strictly speaking, however, the denominator is unknown; there is no way of knowing whether the set of located studies is representative of the full set of existing studies. More elaborate methods, such as estimation of population size using capture-recapture models[7] can be employed.

Precision, the second concept, relates to the false positive rate and is calculated as

$$\text{Precision} = \frac{\text{Number retrieved and relevant}}{\text{Number retrieved}} \times 100.$$

The goal is to have high recall and high precision literature retrievals. Increasing recall and precision can be accomplished by utilizing multiple methods of search strategies. Regardless of how successful the meta-analyst feels the search was, the meta-analyst needs to assess the presence and impact of publication bias on subsequent inferences (see Diagnostics in Section 4.4).

2.4. Methods for searching the literature

A manual search of databases requires specification of a *search statement* and a method of searching. Most research libraries add *controlled vocabulary* terms to the bibliographic record. For example, classification codes and subject headings for books and descriptors for articles comprise controlled vocabulary. In contrast, *natural language* terms are those that may appear in the title or abstract and are not assigned by the research library. Thus, it is important to combine the two sets of terms when defining the search statement, for example

Select State of the Art Review AND Stroke AND Randomized-Controlled-Trial.

There are also different methods of searching the literature. A backward search involves identifying a publication and then moving to earlier items in the citation. A forward search identifies a publication and then searches all items that later cite the publication. The two methods could

yield different publications. The selection principles of the search, such as the scope, including the subject headings, language constraints and the domain, including sources that failed to yield studies, should be documented throughout the data synthesis and subsequently reported.

3. EVALUATING THE RETRIEVED LITERATURE

Once the literature has been retrieved, then the study results need to be *coded* into a database and the criteria used in accepting or rejecting a study to be meta-analysed need to be decided upon. Given the vast quantities of heterogeneous literature, the first task is more daunting than it may first appear. For example, what information should be collected and coded, how many coders should be utilized, how should the coders be trained, and how should they be assessed? The type of items that should be collected should include the characteristics of: the report (author, year, source of publication); the study (scope of the sample, types of populations, and overall characteristics of the study such as the sociodemographic level of the study population); the patients (demographic and clinical features of the study participants); the research design and features (observational or randomized, sampling mechanism, treatment assignment mechanism, compliance rates, attrition rates, type of survey, non-response rates); the treatment (duration, dose, timing, mode of delivery); and the effect size (sample size, nature of outcome, estimate and standard error). The amount and quality of information in any report will depend on several factors not limited to the author's professional training and experience as well as his/her writing capabilities and the journal. A well-designed study to determine the quality of the coding process should also be implemented; see the article by Sands and Murphy[8] describing the inter- and intra-coder reliability in synthesizing the literature.

Rules for assessing the *quality* of the studies and determining their relevance to the objective of the meta-analysis are not clear. The quality and relevance of a study hinges upon the current state-of-the-art. To this end, Chapter 2 of Cook and Campbell[9] provides a framework describing several features of a study, that if threatened, impact on the interpretation and quality of a primary study. For example, suppose that there is bias in who received lidocaine and who did not in study 1 (see Table I) in such a way that sicker patients were given the control treatment. If this randomization bias occurred, then one would expect mortality differences to be smaller between the treatment and control groups if lidocaine was truly detrimental. Such a bias would impact the internal validity of the study. Alternatively, suppose that the power to detect a difference between the treatment and control group is low in study 1; this fact may threaten the statistical validity of the study. Threats of the extent to which one can generalize to other groups of people or settings compromise the external validity of a study. For example, can a relationship observed in a private hospital be generalized to all hospitals?

The approach by Chalmers and his colleagues[10] is similar to the framework proposed by Cook and Campbell but is restricted to focusing on randomized trials. Three areas are assessed in their framework: study design; implementation, and analysis. Two readers blinded to the authors, source, results, and discussion of the primary study make ratings regarding the study's quality. The end result is a percentage score that may be incorporated into a sensitivity analysis at the analytic stage. For example, the meta-analyst may want to examine how overall inferences change if the analysis is restricted to high quality studies.

The validity framework proposed by Cook and Campbell as well as the scoring methods proposed by Chalmers *et al.* provide a set of criteria for making decisions regarding the inclusion/exclusion of studies. A formal approach to deciding the ultimate inclusion status of a study may

be undertaken using a panel of judges/experts. For example, in the lidocaine meta-analysis:

> *The studies were accepted or rejected by two reviewers. A third review was conducted with blinded 'Methods' and 'Results' sections, and there was agreement on the acceptability of all studies included and excluded*,[1] *p.* 2694.

Note that both methods of quality assessment provide a systematic approach to describe primary studies, to explain heterogeneity, and to assess sensitivity of results.

4. COMBINING THE STUDIES

Once the primary studies have been collected and coded, the meta-analyst needs to identify a summary measure common to all studies and subsequently combine the measure. A review of summary measures based on discrete data, such as risk differences, and those based on continuous data, such as standardized mean differences, follows. The fixed-effects and random-effects models are formalized and inferential methods in each are presented. The section concludes with an introduction to model diagnostics.

4.1. Defining the study outcome

Often the meta-analyst has little control over the choice of the summary measure because most of the decision is dictated by what was employed in the primary studies. For example, if risk differences are reported in the primary studies rather than survival times, then the analyst has little choice but to utilize the average risk difference as the summary statistic in the meta-analysis. In many settings, however, different summary measures will be reported across the primary studies. Success may be defined as 30-day survival in one study, as symptom-free survival in another, and as in-hospital survival in yet another. It now becomes the job of the analyst to create a summary statistic that is comparable across all the studies. In some situations this task will be impossible. In this section, three classes of outcome measures are described: measures based on discrete outcome data such as differences in proportions; those based on continuous data that may generally be thought of as means, and a miscellaneous set of outcome measures that may be based on test statistics. The three classes are not exhaustive but are meant to introduce the reader to the challenges involved in creating a summary statistic comparable across studies.

4.1.1. Risk differences, relative risks and odds ratios

Table III demonstrates three potential study summary statistics for binary measurements using the lidocaine data: the difference between two probabilities (risk difference), the ratio of two probabilities (relative risk), and the ratio of the odds for the treated group to the odds for the control group (odds ratio). Risk differences are easy to interpret, are defined for boundary values (proportions of 0 or 1), and are approximately normally distributed for modest sample sizes. Relative risks and odds ratios are typically analysed on the logarithmic scale, but, unlike the risk difference, are not defined for boundary values.

Generally inferences in the lidocaine studies remain the same regardless of the choice of summary statistic. The direction and significance of the study-specific effects are essentially the same regardless of the summary statistic selected. Each 95 per cent confidence interval for the risk difference covers 0 and similarly, the relative risks and odds ratios cover 1. For example, in study 4 with a total sample size of 213 patients, an excess of 1·8 per cent of patients treated with lidocaine

Table III. Study summaries for the lidocaine example. T denotes treatment group, C denotes control group, $q_i = 1 - p_i$; n_{Ti} and n_{Ci} denote the total number of treated and control patients, respectively; and a, b, c, and d denote the number of observations in each of the cells defined by the treatment (lidocaine or control) and outcome (dead or alive) table. The confidence intervals for the relative risk and odds ratio are computed on the logarithmic scale and transformed back to the original scale

	Risk difference	Relative risk	Odds ratio
Parameter	$D = P_T - P_C$	$R = P_T/P_C$	$\Omega = \frac{P_T/(1-P_T)}{P_C/(1-P_C)}$
Estimator	$d_i = \hat{p}_{Ti} - \hat{p}_{Ci}$	$r_i = \frac{\hat{p}_{Ti}}{\hat{p}_{Ci}}$	$\omega_i = \frac{\hat{p}_{Ti}\hat{q}_{Ci}}{\hat{q}_{Ti}\hat{p}_{Ci}}$
Standard error	$s_{d_i} = \sqrt{\left(\frac{p_{Ti}q_{Ti}}{n_{Ti}} + \frac{p_{Ci}q_{Ci}}{n_{Ci}}\right)}$	$s_{\mathrm{Log}(r_i)} = \sqrt{\left(\frac{q_{Ti}}{n_{Ti}p_{Ti}} + \frac{q_{Ci}}{n_{Ci}p_{Ci}}\right)}$	$s_{\mathrm{Log}(\omega_i)} = \sqrt{\left(\frac{1}{n_a} + \frac{1}{n_b} + \frac{1}{n_c} + \frac{1}{n_d}\right)}$

Study	Sample size	d_i (%)	95% CI	r_i	95% CI	ω_i	95% CI
1	82	2·8	(−5·5, 11·1)	2·2	(0·2, 23·4)	2·3	(0·2, 26·1)
2	88	0·0	(−12·0, 12·0)	1·0	(0·3, 3·8)	1·0	(0·2, 4·3)
3	217	2·0	(−3·6, 7·6)	1·5	(0·5, 5·3)	1·6	(0·4, 5·7)
4	213	1·8	(−4·7, 8·3)	1·4	(0·4, 4·1)	1·4	(0·4, 4·5)
5	216	3·5	(−2·0, 9·1)	2·2	(0·6, 8·5)	2·3	(0·6, 9·3)
6	300	4·4	(−0·5, 9·3)	2·6	(0·8, 8·0)	2·7	(0·8, 8·8)

died compared to the control patients with 95 per cent CI for the underlying (unknown) risk difference ranging from – 4·7 per cent to 8·3 per cent. The relative risk and odds ratio estimates are both 1·4, implying that a treated patient is 1·4 times as likely as an untreated patient to die, but both confidence intervals cover 1·0. To compute the approximate 95 per cent confidence interval for the log relative risk corresponding to the ith study, the lower and upper limits are calculated using $\log(r_i) \pm 1{\cdot}96 s_i$ where s_i is the standard error of the log relative-risk, as $(\frac{1-p_{Ti}}{n_{Ti}p_{Ti}} + \frac{1-p_{Ci}}{n_{Ci}p_{Ci}})^{\frac{1}{2}}$, with p and n denoting the mortality rate and sample size, respectively. In the case of study 4, the lower and upper 95 per cent limits are

$$\begin{aligned}(\text{lower}, \text{upper}) &= \log\left(\frac{0{\cdot}068}{0{\cdot}050}\right) \pm 1{\cdot}96\left(\frac{0{\cdot}932}{103 \times 0{\cdot}068} + \frac{0{\cdot}950}{100 \times 0{\cdot}005}\right)^{1/2} \\ &= 0{\cdot}3075 \pm 1{\cdot}96(0{\cdot}32307)^{1/2} \\ &= (-0{\cdot}81, 1{\cdot}42).\end{aligned}$$

Exponentiating yields a lower limit of $e^{-0{\cdot}81} = 0{\cdot}44$ and an upper limit of $e^{1{\cdot}42} = 4{\cdot}14$ for the relative risk.

4.1.2. Means and effect sizes

When the primary studies report means, $\bar{x}$, in each treatment arm, such as in the stroke data displayed in Table II, the analyst may calculate the mean difference and the associated measure of precision for each study. Let i index study, T the treatment group, C the control group, and

$n_{\mathrm{T}i}$ and $n_{\mathrm{C}i}$, the respective sample sizes in the two arms. A potential summary measure is the difference in means, $Y_i = \bar{x}_{\mathrm{T}i} - \bar{x}_{\mathrm{C}i}$ with standard error, s_i, calculated as

$$s_i^2 = s_{pi}^2 \left(\frac{1}{n_{\mathrm{T}i}} + \frac{1}{n_{\mathrm{C}i}} \right) \quad \text{with} \quad s_{pi}^2 = \frac{(n_{\mathrm{T}i} - 1)s_{\mathrm{T}i}^2 + (n_{\mathrm{C}i} - 1)s_{\mathrm{C}i}^2}{n_{\mathrm{T}i} + n_{\mathrm{C}i} - 2}$$

where $s_{\mathrm{T}i}^2$ and $s_{\mathrm{C}i}^2$ are the treatment and control group sample variances, respectively, for the ith study. Figure 2 displays the study means, Y_i, and 95 per cent intervals based on s_i for the stroke study. In Figure 2, study 3 comprised 146 patients, specialist stroke unit care patients remained in the hospital, on average, 55 days less than patients managed routinely, with 95 per cent confidence intervals for the true difference ranging from – 61 days to – 48 days. In the case when there is no direct measure common to all the studies, it may be possible to transform the study-specific summary to a standardized (scale-free) statistic denoted an *effect size*. One common estimator of effect size is the *standardized mean difference* which is calculated as the difference of means divided by the variability of the measures. For example, using $\mathrm{N}(\mu, \sigma^2)$ to denote normally distributed with mean μ and variance σ^2, if

$$Y_{ij}^{\mathrm{T}} \sim \mathrm{N}(\mu^{\mathrm{T}}, \sigma^2); \quad j = 1, 2, \ldots, n_{\mathrm{T}i}$$
$$Y_{ij}^{\mathrm{C}} \sim \mathrm{N}(\mu^{\mathrm{C}}, \sigma^2); \quad j = 1, 2, \ldots, n_{\mathrm{C}i}$$

then the standardized mean difference is defined as

$$\delta = \frac{\mu^{\mathrm{T}} - \mu^{\mathrm{C}}}{\sigma}.$$

δ represents the gain (or loss) as the fraction of the variability of the measurements. An estimator of δ, denoted *Hedges*' g, is defined as $h_i = \frac{\bar{Y}_i^{\mathrm{T}} - \bar{Y}_i^{\mathrm{C}}}{s_p}$. The consequences of dividing by an estimate of the standard deviation is to have a unitless summary measure so that in instances when 'success' is measured in different ways across the studies, the results from the primary studies can be transformed to unitless measures and then pooled. The estimated variance of h_i is

$$\left(\frac{1}{n_{\mathrm{T}i}} + \frac{1}{n_{\mathrm{C}i}} \right) + \frac{\hat{\delta}^2}{2(n_{\mathrm{T}i} + n_{\mathrm{C}i})},$$

where $\hat{\delta}^2$ is the sample estimate of δ^2.

4.1.3. Other measures

When the summary data from the primary studies consist of test statistics, then it is sometimes possible to recover the estimated effect size if the appropriate pieces of information are also reported. For example, if the z-statistic is reported, the estimated standardized mean difference may be calculated as

$$\hat{\delta} = z \sqrt{\left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}} \right)}.$$

If the study summaries are significance levels (p-values) then these may also be combined (Hedges and Olkin,[4] Chapter 3) although this method adds little insight in terms of the size of the effect and its direction. The reader is referred to Chapter 2 of Rosenthal[5] for a summary of the relationship between effect sizes and tests of significance.

4.2. Modelling variation in meta-analysis

There are at least three sources of variation to consider before combining summary statistics across studies. First, sampling error may vary among studies. For example, sample sizes range from 82 to 300 in the lidocaine example and from 21 to 311 in the stroke example resulting in study summaries estimated with varying degrees of precision. Second, study-level characteristics may differ among the studies. The stroke studies were conducted at both for-profit and not-for-profit hospitals and there may be reason to believe the treatment effect is different in these two hospital types. Third, there may exist inter-study variation. The fixed-effects model introduced in Section 4.2.1 assumes each study is measuring the same underlying parameter and that there is no inter-study variation. Conversely, the random-effects model (introduced in Section 4.2.2) assumes each study is associated with a different but related parameter.

4.2.1. Fixed-effects model

A fixed-effects model assumes that each study summary statistic, Y_i, is a realization from a population of study estimates with common mean θ (Figure 3). Let θ be the central parameter of interest and assume there are $i = 1, 2, \ldots, k$ independent studies. Assume that Y_i is such that $\mathrm{E}(Y_i) = \theta$ and let $s_i^2 = \mathrm{var}(Y_i)$ be the variance of the summary statistic in the ith study. For moderately large study sizes, each Y_i should be asymptotically normally distributed (by the central limit theorem) and approximately unbiased. Thus

$$Y_i \overset{\text{indep.}}{\sim} \mathrm{N}(\theta, s_i^2) \quad \text{for } i = 1, 2, \ldots, k \tag{1}$$

and s_i^2 assumed known. The central parameter of interest is θ which quantifies the average treatment effect.

4.2.2. Random-effects model

The random-effects framework postulates that each study summary statistic, Y_i, is a draw from a distribution with a study-specific mean, θ_i, and variance, s_i^2:

$$Y_i \mid \theta_i, s_i^2 \overset{\text{indep.}}{\sim} \mathrm{N}(\theta_i, s_i^2). \tag{2}$$

Furthermore, each study-specific mean, θ_i, is assumed to be a draw from some superpopulation of effects (see discussion in Section 2.1) with mean θ and variance τ^2 as depicted in Figure 4, with

$$\theta_i \mid \theta, \tau^2 \overset{\text{indep.}}{\sim} \mathrm{N}(\theta, \tau^2). \tag{3}$$

θ and τ^2 are referred to as *hyperparameters* and represent, respectively, the average treatment effect and inter-study variation.

Note that, given the hyperparameters, the distribution of each study summary measure, Y_i, after averaging over the study-specific effects, is Normal with mean θ and variance $s_i^2 + \tau^2$. As in the fixed-effects model, θ is a parameter of central interest; however, the between-study variation, τ^2, plays an important role and must also be estimated. In addition to the average treatment effect, it is also possible to derive estimates of the study-specific effects, θ_i, that are useful for inferences regarding identifying particularly effective studies. The distribution of θ_i, conditional on the observed data and the hyperparameters, denoted the *posterior distribution*, is

$$\theta_i \mid \mathbf{y}, \theta, \tau^2 \sim \mathrm{N}(B_i\theta + (1 - B_i)Y_i, s_i^2(1 - B_i)) \tag{4}$$

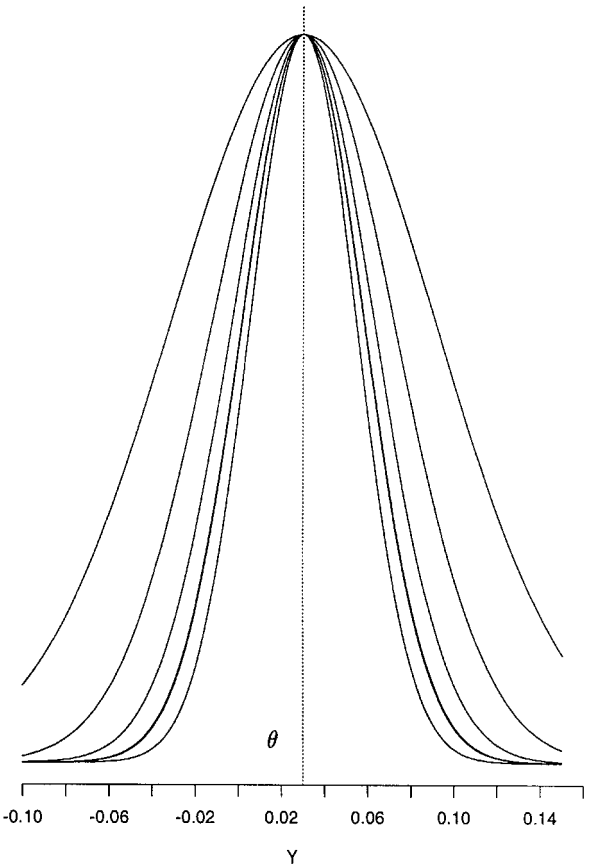

Figure 3. Fixed-effects model. The distribution of five hypothetical study statistics under the assumptions of the fixed-effects model. Each study sample mean, Y_i, provides an estimate of a common mean, θ (denoted by the dashed vertical line). The difference among the five studies rests only on, s_i^2, how well each study sample mean estimates θ

where $\mathbf{y} = (Y_1, Y_2, \ldots, Y_k)$. B_i, defined as $s_i^2/(s_i^2 + \tau^2)$, is commonly referred to as the *shrinkage* factor for the ith study. The larger the inter-study variation, τ^2, the smaller the shrinkage B_i of the observed study effects. Because $0 \leqslant B_i \leqslant 1$, the mean of θ_i in equation (4) is a compromise between the average treatment effect, θ, and the observed study summary statistic, Y_i. When $\tau^2 = 0$, shrinkage is maximized with $B_i = 1$ so that $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$ and the random-effects model corresponds to the fixed-effects model.

4.3. Inference

In order to account for differences in sample sizes and study-level characteristics, studies are stratified and then combined. That is, rather than estimating the true effect of lidocaine as the difference between the total fraction dying in the treatment and control groups, $\frac{37}{557} - \frac{21}{549}$ (Table I),

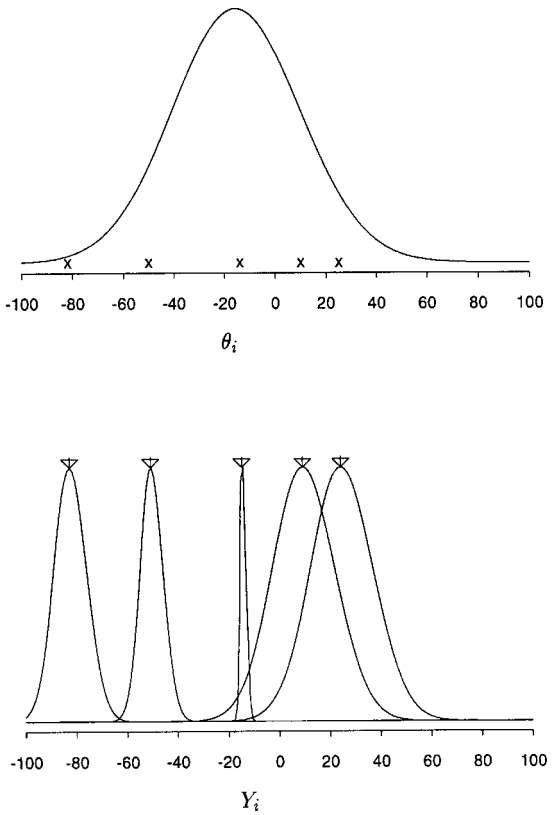

Figure 4. Random-effects model. The distribution of five hypothetical study statistics under the assumptions of the random-effects model. Each effect, θ_i, is drawn from a superpopulation with mean θ and variance τ^2 (upper plot). The study-specific summary statistics, Y_i, are then generated from a distribution with mean determined by θ_i (denoted by $\times$ in the upper plot) and variance s_i^2 (lower plots). In the example, each of the five effects generated the five study results (lower plots)

a weighted average of the estimates from each study is taken. Similarly, a weighted average of the estimates of the treatment effects from each for-profit hospital study and a weighted average of the estimates from the not-for-profit hospital studies could be taken in the stroke example. However, the distinction of whether each study (or each set of studies) measures a common parameter remains. Therefore the convention is to first perform a test of homogeneity of means. If no significant inter-study variation is found, then a fixed-effects approach is adopted; otherwise the meta-analyst either adopts a random-effects approach or identifies study characteristics that stratifies the studies into subsets with homogeneous effects. The test of homogeneity is next described and is followed by a description of inferential modes for a fixed-effects and random-effects model. Maximum likelihood, restricted maximum likelihood and Bayesian methods are given for both types of models and summarized at the conclusion of this section.

4.3.1. Test of homogeneity

The fixed-effects model (equation (1)) assumes that the k study-specific summary statistics share a common mean θ. A statistical test for the homogeneity of study means is equivalent to testing

$$H_0: \quad \theta = \theta_1 = \theta_2 = \cdots = \theta_k \text{ against}$$

$$H_1: \quad \text{At least one } \theta_i \text{ different.}$$

Under H_0, for large sample sizes, $Q_W = \sum_i^k W_i(Y_i - \hat{\theta}_{\text{MLE}})^2 \sim \chi^2_{k-1}$ where $\hat{\theta}_{\text{MLE}} = \sum W_i Y_i / \sum W_i$ and $W_i = 1/s_i^2$. If Q_W is greater than the $100(1-\alpha)$ percentile of the χ^2_{k-1} distribution, then the hypothesis of equal means, H_0, would be rejected at the 100 per cent level. If H_0 is rejected, the meta-analyst may conclude that the study means arose from two or more distinct populations and proceed by either attempting to identify covariates that stratify studies into the homogeneous populations or estimating a random-effects model. If H_0 cannot be rejected the investigator would conclude that the k studies share a common mean, θ, and estimate θ using $\hat{\theta}_{\text{MLE}}$. Tests of homogeneity have low power against the alternative $\text{var}(\theta_i) > 0$. Note that not rejecting H_0 is equivalent to asserting that the amount of between-study variation is small.

4.3.2. Fixed-effects model

When s_i^2 is assumed known, the log-likelihood for θ, $\log(L(\theta \mid \mathbf{y}, \mathbf{s}^2))$ is proportional to $\sum_i(\frac{(Y_i-\theta)^2}{s_i^2})$ leading to the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i} \quad \text{with } W_i = \frac{1}{s_i^2} \tag{5}$$

where $\mathbf{s} = (s_1^2, s_2^2, \ldots, s_k^2)$. Standard inferences about θ are available using the fact that $\hat{\theta}_{\text{MLE}} \sim N(\theta, (\sum_i W_i)^{-1})$. A Bayesian approach may be adopted by specifying a prior distribution for θ, for example, $\theta \sim N(0, \sigma_0^2)$, and calculating the posterior distribution

$$\theta \mid \mathbf{y}, \mathbf{s}, \sigma_0^2 \sim N\left(\left[\sum_i W_i + \sigma_0^{-2}\right]^{-1}\left(\sum_i W_i Y_i\right), \left[\sum_i W_i + \sigma_0^{-2}\right]^{-1}\right).$$

The estimator of θ is the posterior mean

$$\tilde{\theta}_B = \left[\sum_i W_i + \sigma_0^{-2}\right]^{-1}\left(\sum_i W_i Y_i\right). \tag{6}$$

If σ_0^2 is large, then the posterior mean coincides with the maximum likelihood estimator.

4.3.3. Random-effects model

If τ^2 is known then the MLE of θ is given by

$$\hat{\theta}(\tau)_{\text{MLE}} = \frac{\sum_i W_i(\tau) Y_i}{\sum_k W_i(\tau)} \quad \text{with} \quad W_i(\tau) = \frac{1}{s_i^2 + \tau^2}. \tag{7}$$

However, in the more realistic case of unknown τ^2, two common methods of inference can be employed: restricted maximum likelihood (REML) or Bayesian.

Restricted Maximum Likelihood (REML) This is a method for estimating variance components in a general linear model.[11, 12] Using the marginal distribution for $\mathbf{y}$, the log-likelihood to be maximized is

$$\log(L(\theta,\tau^2\,|\,\mathbf{s}^2,\mathbf{y}) \propto \sum_i \left\{ \log(s_i^2+\tau^2) + \frac{(Y_i-\hat{\theta}_{\mathrm{R}})^2}{s_i^2+\tau^2} \right\} + \log\left(\sum (s_i^2+\tau^2)^{-1}\right).$$

The REML of τ^2 is the solution to

$$\hat{\tau}^2_{\mathrm{R}} = \frac{\sum_i w_i^2(\hat{\tau})\left(\frac{k}{k-1}(Y_i-\hat{\theta}_{\mathrm{R}})^2 - s_i^2\right)}{\sum_i w_i^2(\hat{\tau})}.$$

The estimator for the population mean is then calculated as

$$\hat{\theta}_{\mathrm{R}} = \frac{\sum_i^k w_i(\hat{\tau}_{\mathrm{R}})Y_i}{\sum_i^k w_i(\hat{\tau}_{\mathrm{R}})}; \quad w_i(\hat{\tau}_{\mathrm{R}}) = \frac{1}{s_i^2+\hat{\tau}^2_{\mathrm{R}}} \tag{8}$$

and inferences are made using $\hat{\theta}_{\mathrm{R}} \sim \mathrm{N}(\theta, (\sum_i w_i(\hat{\tau}_{\mathrm{R}}))^{-1})$. An estimator for θ_i can be calculated by substituting the REML estimates for the hyperparameters in equation (4). This type of approximation to the posterior distribution is known as *empirical Bayes* and results in $\hat{\theta}_i^{\mathrm{R}} = (1-\hat{B}_i^{\mathrm{R}})Y_i + \hat{B}_i^{\mathrm{R}}\hat{\theta}_{\mathrm{R}}$ where $\hat{B}_i^{\mathrm{R}} = \frac{s_i^2}{s_i^2+\hat{\tau}^2_{\mathrm{R}}}$ is the shrinkage estimate. Inferences for the study-specific effects are made using $\hat{\theta}_i^{\mathrm{R}} \sim \mathrm{N}(\theta_i, s_i^2(1-\hat{B}_i^{\mathrm{R}}))$. Models can be estimated using the SAS procedure Proc Mixed (see Section 5). Note that the empirical Bayes approximation is deficient in that it ignores the uncertainty in the hyperparameters, $\{\theta,\tau^2\}$.

Fully Bayesian In order to reflect the uncertainty in the estimates of hyperparameters θ and τ^2 (equation (3)), a fully Bayesian approach can be adopted.[13–16] Prior distributions on the unknown parameters are specified and inferences about the population effect θ (and the θ_is) can be made by integrating out the unknown parameters over the joint posterior distribution of all the parameters. Let $\theta \sim \mathrm{N}(0,a^2)$ and $\tau^{-2} \sim \mathrm{gamma}(c,d)$ with $E(\tau^{-2}) = c/d$ and $\mathrm{var}(\tau^{-2}) = c/d^2$. Then the joint posterior distribution for $V = \{\theta,\theta_1,\ldots,\theta_k,\tau^2\}$ is calculated as:

$$p(V\,|\,\mathbf{y},\mathbf{s^2}) \propto \prod_i p(\theta_i\,|\,y_i,s_i^2)\,p(\theta_i\,|\,\theta,\tau^2)\,p(\theta)\,p(\tau^2).$$

Inferences are conducted using summaries of the posterior distribution, for example

$$\hat{\theta}_{\mathrm{B}} = E(\theta\,|\,\mathbf{y},\mathbf{s}^2) = \int_\theta \theta \int_{\theta_i,\tau^2} \{p(V)\,\mathrm{d}\theta_i\,\mathrm{d}\tau^2\}\,\mathrm{d}\theta. \tag{9}$$

The integral in equation (9) may be analytically tractable when the prior and likelihood are conjugate. Typically, though, the integral must be evaluated numerically. In cases such as these, Monte Carlo approximations to the posterior, such as those employed in BUGS (see Section 5), may be utilized. Other approximations to the posterior distributions are also available. For example, Morris[17] and Morris and Normand[14] proposed an approximation to mean of the posterior

distribution for τ^2, denoted the *adjusted likelihood estimator* derived as a result of applying a Pearson approximation to the posterior density for τ^2.

Method of Moments (MOM) A third estimator of τ^2 is provided by the homogeneity test. By equating Q_W with its corresponding expected value, DerSimonian and Laird[18] proposed a non-iterative (method of moments) estimator of τ^2 defined as

$$\hat{\tau}^2_{\mathrm{DL}} = \max\left\{0, \frac{Q_W - (k-1)}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}}\right\}.$$

This leads to

$$\hat{\theta}_{\mathrm{DL}} = \frac{\sum_i w_i(\hat{\tau}_{\mathrm{DL}})Y_i}{\sum_i w_i(\hat{\tau}_{\mathrm{DL}})} \quad \text{with } w_i(\hat{\tau}_{\mathrm{DL}}) = \frac{1}{s_i^2 + \hat{\tau}^2_{\mathrm{DL}}}. \tag{10}$$

$\hat{\theta}_{\mathrm{DL}}$ is also denoted Cochran's semi-weighted estimator of θ and can be easily programmed using most software packages. Table IV contains a summary of the estimators that we presented above.

4.4. Diagnostics

Once the data have been collected and analyzed the meta-analyst needs to assess the appropriateness of the assumptions that have been made. Two aspects of diagnostics are discussed. A systematic approach to investigating how sensitive the results are to the method of analysis or to changes in the data, denoted a *sensitivity analysis*, is next introduced. Methods for assessing and adjusting the meta-analysis when there is a biased sampling mechanism are also presented.

Note that prior to the analysis of the set of primary studies, several quantitative summaries will have already been collected. The success of the literature retrieval as measured by the recall and precision or as estimated using capture-recapture models give an indication as to the representativeness of the collected literature (Section 2.3). If a panel of raters has been used to determine the appropriateness of the studies, then inter- and intra-rater reliability statistics are useful measures of quality; similarly, inter- and intra-coder reliability statistics provide guidance as to the accuracy of the data underlying the meta-analysis.

4.4.1. Sensitivity analysis

An exploratory analysis of the primary data, for example, the study-specific estimates, should first be undertaken in order to understand important features of the data. For example, a box plot of the study effects will indicate typical values, spread (skewness, multi-modal etc.), and tails (presence of outliers). The box plots may be stratified by characteristics of the studies (including quality scores if available) in order to understand how and why studies differ. However, if the number of studies is small, as in the two examples in this article, then the meta-analyst is limited to the range of descriptive analyses that can be undertaken.

The meta-analyst should estimate *both* a fixed-effects and a random-effects model and compare the results of both. Sensitivity to the distributional assumptions can be assessed by assuming different distributions for the study effects and comparing subsequent inferences. For example, the analyst may assume that the underlying study effects, θ_i, arise from a Student-t distribution, thereby permitting heavier tails than those arising from a Normal distribution. Moreover, within a model,

Table IV. Summary of estimators for fixed-effects and random-effects models. Observed Fisher information denotes the matrix inverse of the second derivatives of the log-likelihood function evaluated at the REML estimates. $V = \{\theta, \theta_1, \ldots, \theta_k, \tau^2\}$ and $g(V)$ denotes a function of the parameters, for example, $P(\theta_i > 0)$

Method	Parameter	Estimator	Variance
Fixed-effects model: $Y_i \sim \mathrm{N}(\theta, s_i^2)$			
MLE	θ	$\hat{\theta}_{\mathrm{MLE}} = \dfrac{\sum_i W_i Y_i}{\sum_i W_i}$	$(\sum_i W_i)^{-1}$
		$W_i = 1/s_i^2$ assumed known	
Bayesian	θ	$\tilde{\theta}_{\mathrm{B}} = [\sum_i W_i + \sigma_0^{-2}]^{-1}(\sum_i W_i Y_i)$	$[\sum_i W_i + \sigma_0^{-2}]^{-1}$
		$W_i = 1/s_i^2$, σ_0^2 assumed known	
Random-effects model $Y_i \mid \theta_i \sim \mathrm{N}(\theta_i, s_i^2)$; $\theta_i \mid \theta, \tau^2 \sim \mathrm{N}(\theta, \tau^2)$			
DerSimonian and Laird	τ^2	$\hat{\tau}^2_{\mathrm{DL}} = \max\left\{0, \dfrac{Q_W - (k-1)}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}}\right\}$	None proposed
(Method of moments)	θ	$\hat{\theta}_{\mathrm{DL}} = \dfrac{\sum_i w_i(\hat{\tau}_{\mathrm{DL}}) Y_i}{\sum_i w_i(\hat{\tau}_{\mathrm{DL}})}$	$(\sum_i w_i(\hat{\tau}_{\mathrm{DL}}))^{-1}$
		$W_i = 1/s_i^2$, $w_i(\hat{\tau}_{\mathrm{DL}}) = \frac{1}{s_i^2 + \hat{\tau}^2_{\mathrm{DL}}}$ assumed known	
REML	τ^2	$\hat{\tau}^2_{\mathrm{R}} = \dfrac{\sum_i w_i^2(\hat{\tau})(\frac{k}{k-1}(Y_i - \hat{\theta}_{\mathrm{R}})^2 - s_i^2)}{\sum_i w_i^2(\hat{\tau})}$	Observed Fisher information
	θ	$\hat{\theta}_{\mathrm{R}} = \dfrac{\sum_i w_i(\hat{\tau}_{\mathrm{R}}) Y_i}{\sum_i w_i(\hat{\tau}_{\mathrm{R}})}$	$(\sum_i w_i(\hat{\tau}_{\mathrm{R}}))^{-1}$
Empirical Bayes	θ_i	$\hat{\theta}_i^{\mathrm{R}} = \hat{B}_i^{\mathrm{R}} \hat{\theta}_{\mathrm{R}} + (1 - \hat{B}_i^{\mathrm{R}}) Y_i$	$s_i^2(1 - \hat{B}_i{}^{\mathrm{R}})$
		$w_i(\hat{\tau}_{\mathrm{R}}) = \frac{1}{s_i^2 + \hat{\tau}^2_{\mathrm{R}}}, \hat{B}_i^{\mathrm{R}} = \frac{s_i^2}{s_i^2 + \hat{\tau}^2_{\mathrm{R}}}$ assumed known	
Bayesian	τ^2	$\hat{\tau}^2_{\mathrm{B}} = \int \tau^2 \hat{p}(V \mid \mathbf{y}, \mathbf{s}) \mathrm{d}\theta_i \, \mathrm{d}\theta \, \mathrm{d}\tau^2$	From empirical distribution
	θ	$\hat{\theta}_{\mathrm{B}} = \int \theta \hat{p}(V \mid \mathbf{y}, \mathbf{s}) \mathrm{d}\theta_i \, \mathrm{d}\tau^2 \, \mathrm{d}\theta$	From empirical distribution
	θ_i	$\hat{\theta}_i^{\mathrm{B}} = \int \theta_i \hat{p}(V \mid \mathbf{y}, \mathbf{s}) \mathrm{d}\theta_j \, \mathrm{d}\theta \, \mathrm{d}\tau^2 \, \mathrm{d}\theta_i$	From empirical distribution
	g(V)	$\hat{g}(V) = \int g(V) \hat{p}(V \mid \mathbf{y}, \mathbf{s}) \mathrm{d}V$	From empirical distribution
		Prior distribution for hyperparameters assumed known	

the meta-analyst should determine how sensitive the combined estimate is to any one study or group of studies. This can be accomplished by leaving one study out, calculating the combined effect of the remaining studies, and comparing the results with the combined effect based on all the studies.

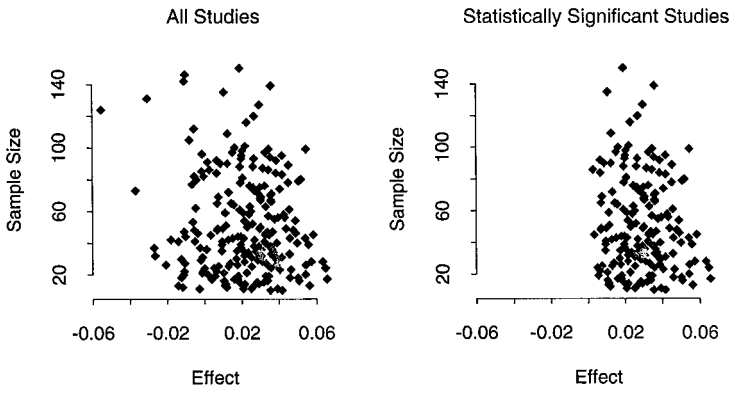

Figure 5. Two funnel plots based on simulated data. 230 effects simulated assuming $\theta_i \sim \mathrm{N}(0{\cdot}02, (0{\cdot}02)^2)$; $y_i \mid \theta_i, n_i \sim \mathrm{N}(\theta_i, (0{\cdot}01)^2/n_i)$ with n_i ranging from 10 to 150. The rightmost plot displays those studies statistically significant at 0·05 level

4.4.2. Publication bias

As indicated in previous sections, a major concern of the meta-analyst is whether the collection of analysed studies were selected using a biased mechanism. One may conjecture several study features that may be correlated with the propensity with which a study appears in the published literature. One such feature is the significance level associated with the statistical test used in the primary study. For example, journals are typically more likely to publish results that establish a difference than those that do not; moreover, even if a non-significant result is published, there is a tendency to publish few statistical details of such an (apparently) uninteresting result. With this in mind, a graphic device known as a *funnel plot*[19] can be employed to detect the presence of publication bias. A funnel plot is a scatter plot of sample size or other measure of precision on the y-axis versus the estimated effect size on the x-axis for a group of studies. Because smaller studies, such as phase I studies, will demonstrate more variability among effects and will be more prevalent than larger studies, then the plot should look like a funnel. Publication bias is suspected if there is a 'piece' missing from the plot.

Figure 5 demonstrates, using 230 simulated study results, the impact on the funnel plot when there is bias against publishing non-significant results. The leftmost plot displays the simulated summaries for all the studies while the rightmost plot displays the simulated summaries for studies that are statistically significant at the 0·05 level. Note that a section of the funnel in the lower portion of the plot is missing. Unfortunately, detecting bias via a funnel plot is not as obvious as it is in this contrived example. There may be several types of biasing mechanism present at any given time; for example, there may be both a bias in publishing results from small studies (even if significant) as well as against publishing non-significant results.

If publication bias is suspected, the meta-analyst may model the selection process into the model in order to correct for the bias.[20–23] One possibility is to view the selection problem as a missing data problem and assume that the studies are missing with probabilities that are a function of their lack of statistical significance.[23] For example, the second plot in Figure 5 was created using

$$p_i(z) = \begin{cases} 0 & \text{if } z \leqslant 1{\cdot}96 \\ 1 & \text{if } z > 1{\cdot}96 \end{cases}$$

where $p_i(z)$ denotes the probability of publication of the ith study which depends on its z-value. These proposed methods can be used to provide a broad indication of whether selection bias is present, and if it is, the impact of the bias on estimation.

5. SOFTWARE

Estimation of fixed-effects models in the context of a meta-analysis is generally straightforward and can be coded using standard software packages such as S-plus.[24] Moreover, there are several PC-based packages available to perform such analyses (see for example the review by Normand[25]). Software for estimating random-effects models are generally not as accessible. This is particularly true in the case of meta-analysis because the user will typically want to force the package to make use of the *known* variances, s_i^2, in the estimation process. For these reasons, two packages for performing inference in a random-effects model are next introduced. The first package is the Statistical Analysis System (SAS) and provides estimates using restricted maximum likelihood estimation. The second package is a more recently developed system called BUGS (Bayesian Inference Using Gibbs Sampling) and performs inference within a fully Bayesian framework. This list of software is not exhaustive. For example, DuMouchel[26] has developed an S-plus function to implement a fully Bayesian meta-analysis and is publicly available. Rather, the description that follows is meant to acquaint the meta-analyst with the computational tactics (and tricks) utilized when performing inference in the random-effects model.

5.1. SAS: Proc Mixed

Version 6·11 of SAS provides a procedure denoted Proc Mixed that fits mixed linear models including variance component models. A restatement of the random-effects model in terms of a variance components model is given by the following:

$$\mathbf{Y} = \mu\mathbf{1}_k + \boldsymbol{\delta} + \mathbf{e}; \quad \boldsymbol{\delta} \sim \mathrm{N}_k(\mathbf{0}, \tau^2 I_k)$$

$$\mathbf{e} \sim \mathrm{N}_k(\mathbf{0}, R) \quad \text{where } R = \mathrm{diag}(s_1^2, s_2^2, \ldots, s_k^2)$$

where $\mathbf{Y}$ is a k-vector of observations from the primary studies, $\mathbf{1}_k$ is a k-vector of ones, and I_k is the $k \times k$ identity matrix. It follows that the study-specific effect for study i is $\theta_i = \mu + \delta_i$. The general model estimated in Proc Mixed is

$$Y = X\beta + Zv + e;$$

$$v \sim \mathrm{N}(0, G); \quad e \sim \mathrm{N}(0, R)$$

where β is the fixed effect, v is the random effect, and e is the error at the study level. The procedure permits many different parameterizations for R and G including fixing G and estimating the sampling variance, R, but not vice-versa. Unfortunately, in the meta-analysis problem, the user typically wants to fix the sampling variance R and estimate G. However, users may still use Proc Mixed by reversing the roles of the within-study and between-study specifications and then post-processing the SAS output.


```
filename temp 'los.data';
data stroke;                                    #SET UP DATA TO BE USED IN PROCEDURE:
 infile temp;
 input ntreat lostreat sdtreat ncontrol loscont sdcont;
 diff = lostreat - loscont;
 sp = ((ntreat-1)*sdtreat*sdtreat + (ncontrol - 1)*sdcont*sdcont)/(ntreat + ncontrol - 2);
 vdiff = sp*(1/ntreat + 1/ncontrol);
 styid = _n_;
 row = _n_; col = _n_; value = vdiff;           #SPECIFYING G MATRIX SO THAT THE (ROW,COL) ELEMENT
                                                #IS EQUAL TO 'VALUE'; ALL OTHER ELEMENTS EQUAL 0.
 proc print;
  title 'Stroke Study';

proc mixed data=stroke order=data;              #CALL SAS PROCEDURE; CLASSIFICATION VARIABLE IS SORTED
                                                #BY THE ORDER OF THE INPUT DATA SET.
 class styid;                                   #IDENTIFIES 'STYID' AS A CLASSIFICATION VARIABLE.
 model diff = /p s;                             #MODEL IS INTERCEPT ONLY; PRINT PREDICTED VALUE 'P'
                                                #AND SOLUTION FOR FIXED EFFECTS 'S'.
 random styid / gdata = stroke s;               #G MATRIX IN DATASET 'STROKE'; PRINT SOLUTION
                                                #FOR RANDOM EFFECTS 'S'.
 repeated diag;                                 #SAMPLING VARIANCE MATRIX 'R' IS DIAGONAL.

 make 'Predicted' out=predv;                    #MAKE A SAS DATA SET OF PREDICTED VALUES
 make 'SolutionR' out=randv;                    #MAKE A SAS DATA SET OF RANDOM EFFECTS

data stroke;
 merge stroke predv(keep = pred obser se_pred) randv(keep = est);
 thetai = obser - est; drop obser est;          #CORRECT PREDICTED VALUES TO GET SHRINKAGE
 p_gt_0 = probnorm(thetai/se_pred);             #ESTIMATES AND CALCULATE PROBABILITY THAT
                                                #STUDY-SPECIFIC EFFECT IS GREATER THAN 0
proc print;

run;
endsas;
```

Figure 6. SAS model statements in Proc Mixed. Call to procedure to estimate a variance components model in order to perform a random-effects meta-analysis. Comments are denoted by a # symbol

5.1.1. An Example

Figure 6 displays the SAS statements used in analysing the stroke data. The data is read in from the file 'los.data', the study outcomes are defined (diff) as well as the variance of the study outcomes (vdiff) in the first data step. Because the procedure does not permit fixing G, in order to specify information regarding the known sampling variances, the user must reverse the roles of G and R. In other words, let $G = \mathrm{diag}(s_1^2, s_2^2, \ldots, s_k^2)$ and $R = \tau^2 I_k$. The diagonal elements of G are next specified in the data step using the keywords `row,col` and `value`. This is the sparse representation of G by which the user specifies the value of the entries of the G matrix using the row and column locations (row, col); all other unspecified locations are assumed to be 0. Note

that rather than fitting the model

$$Y_i = \mu + \delta_i + e_i; \quad \delta_i \sim \mathrm{N}(0, \tau^2); \quad e_i \sim \mathrm{N}(0, s_i^2) \tag{11}$$

the model SAS estimates is

$$Y_i = \mu + \delta_i^* + e_i^*; \quad \delta_i^* \sim \mathrm{N}(0, s_i); \quad e_i^* \sim \mathrm{N}(0, \tau^2). \tag{12}$$

The data set is printed and next the procedure `Proc Mixed` is called using the data set stroke and sorted by the order in which they appear in the data set. `Class` designates which variables will be used as stratifiers or classifiers in the subsequent analyses. The `model` statement indicates that the dependent variable is diff and is a function of an intercept term only (note the default is to include an intercept term). The user specifies the '`p`' option to print the predicted values, for example, $\mu + \delta_i^*$, and to print an estimate of the overall mean, μ, using the '`s`' option. The `random` statement designates the random-effects and does not include an intercept by default. Thus, '`styid`' is specified as a random-effect and G is fixed by using the `gdata` option. This option indicates that the G matrix is to read in from the SAS data set '`stroke`' and using the keywords col, row, and value. The '`s`' option in the random statement requests that the estimated random-effects, δ_i^* be printed. The `repeated` statement is used to specify the R matrix. In the example, R is specified using the keyword '`diag`' to indicate that R is a diagonal matrix, $R = \tau^2 I_k$. By default, REML is the estimation method for the covariance parameters.

Post-Processing SAS Output to Get Shrinkage Estimates. Because the roles of G and R have been reversed, in order to obtain the correct study-specific effect estimates, θ_i, the user needs to process the output further. Specifically, the shrinkage estimate for the ith study is calculated as

$$\theta_i = \mu + \delta_i = Y_i - \delta_i^*.$$

In order to calculate this, the user must first create SAS data sets containing the components listed in equation (13). The `make` statements request that an SAS data set containing the predicted values ($\mu + \delta_i^*$ and its standard error) and a data set comprising the random effects (δ_i^* and corresponding standard error) be created. The final data step merges the original data with the predicted values and random effects. The study-specific estimates (denoted `thetai`) are then calculated as described in equation (13) above and the probability that the study-specific estimates are greater than 0 is computed using the `probnorm` function in SAS. Although the SAS Predicted values need to be manipulated, the standard errors corresponding to the estimates are correct.

Figure 7 displays partial output from the call to `Proc Mixed`. The data set stroke and the full SAS output are printed in the Appendix. Three iterations were used before convergence was reached. The estimate of inter-study variation, τ^2, is 685·09 and is printed under the `Covariance Parameter Estimation` along with an estimate of the standard error and a Wald test. The p-value of 0·0713 associated with the Wald test suggests non-homogeneous study means. The estimates of μ and $\{\delta_i^*,\ i = 1, 2, \ldots, 9\}$ are printed under the `Solution for Fixed Effects` and `Solution for Random Effects` sections, respectively. The combined estimate of difference in length of stay, $\hat{\mu}$, is -15 days with a standard error of 9 days.

To examine the shrinkage estimates for any particular study, the meta-analyst refers to the post-processed values under the column labelled THETAI. For example, the estimate corresponding to Study 4 is defined as $\mu + \delta_4$ and is estimated to be -54 days. The probability that this effect is positive is 3 per cent (0·02986); alternatively, there is strong evidence (97 per cent) that specialty

Stroke Study 16:16 Wednesday, October 2, 1996 1

The MIXED Procedure

Class Level Information

Class	Levels	Values
STYID	9	1 2 3 4 5 6 7 8 9

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	63.77953195	
1	3	63.33581077	0.00054886
2	1	63.31711993	0.00000851
3	1	63.31684770	0.00000000

Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm	Estimate	Std Error	Z	Pr > |Z|
DIAG	685.09397633	379.91352734	1.80	0.0713
Residual	1.05643873	.	.	.

Model Fitting Information for DIFF

Description	Value
Observations	9.0000
Variance Estimate	1.0564
Standard Deviation Estimate	1.0278
REML Log Likelihood	-39.0099
Akaike's Information Criterion	-40.0099
Schwarz's Bayesian Criterion	-40.0497
-2 REML Log Likelihood	78.0199

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
INTERCEPT	-15.12158889	8.95356233	8	-1.69	0.1297

Solution for Random Effects

Parameter	Estimate	SE Pred	DDF	T	Pr > |T|
STYID 1	-0.27284292	6.21024468	0	-0.04	.
STYID 2	0.03908618	1.42879248	0	0.03	.
STYID 3	-0.87007274	3.87112874	0	-0.22	.
STYID 4	-10.04910369	11.21602341	0	-0.90	.
STYID 5	0.31841048	4.43620482	0	0.07	.
STYID 6	0.02874070	1.10526162	0	0.03	.
STYID 7	3.19213098	9.21507235	0	0.35	.
STYID 8	0.05934993	2.81953230	0	0.02	.
STYID 9	0.64860109	4.48951554	0	0.14	.

Figure 7. Output from Proc Mixed in SAS

Stroke Study 16:16 Wednesday, October 2, 1996 1

Predicted Values

Observed	Predicted	Var Pred	SE Pred	L95M	U95M	Residual
-20.0000	-15.3944	109.7662	10.4769	.	.	-4.6056
-2.0000	-15.0825	81.7301	9.0405	.	.	13.0825
-55.0000	-15.9917	91.6538	9.5736	.	.	-39.0083
-71.0000	-25.1707	177.1315	13.3091	.	.	-45.8293
-4.0000	-14.8032	95.2559	9.7599	.	.	10.8032
1.0000	-15.0928	81.1020	9.0057	.	.	16.0928
11.0000	-11.9295	145.4908	12.0620	.	.	22.9295
-10.0000	-15.0622	86.2581	9.2875	.	.	5.0622
7.0000	-14.4730	95.6211	9.7786	.	.	21.4730

OBS	NTREAT	LOSTREAT	SDTREAT	NCONTROL	LOSCONT	SDCONT	DIFF	SP
1	155	55	47	156	75	64	-20	3155.55
2	31	27	7	32	29	4	-2	32.23
3	75	64	17	71	119	29	-55	557.33
4	18	66	20	18	137	48	-71	1352.00
5	8	14	8	13	18	11	-4	100.00
6	57	19	7	52	18	4	1	33.27
7	34	52	45	33	41	34	11	1597.18
8	110	21	16	183	31	27	-10	551.83
9	60	30	27	52	23	20	7	576.46

OBS	VDIFF	STYID	ROW	COL	VALUE	THETAI	SE_PRED	P_GT_0
1	40.586	1	1	1	40.586	-19.7272	10.4769	0.02986
2	2.047	2	2	2	2.047	-2.0391	9.0405	0.41078
3	15.281	3	3	3	15.281	-54.1299	9.5736	0.00000
4	150.222	4	4	4	150.222	-60.9509	13.3091	0.00000
5	20.192	5	5	5	20.192	-4.3184	9.7599	0.32908
6	1.224	6	6	6	1.224	0.9713	9.0057	0.54294
7	95.376	7	7	7	95.376	7.8079	12.0620	0.74129
8	8.032	8	8	8	8.032	-10.0593	9.2875	0.13938
9	20.694	9	9	9	20.694	6.3514	9.7786	0.74200

Figure 7. Continued

stroke care is associated with shorter length of stay compared to routine management. Examination of the shrinkage estimate for study 6 indicates that shorter lengths of stays are equally probable when managed under routine management or specialty care ($P\_GT\_0 = 0{\cdot}54$).

5.2. BUGS

BUGS (version 0·501) is a software package written in Modula-2 and distributed as compiled code.[27, 28] The software conducts Bayesian inference using a Monte Carlo Markov chain technique called Gibbs sampling. The basic sampling approach employed in BUGS is adaptive rejection sampling using log-concave distributions. The syntax is surprisingly similar to S-plus so many users will feel comfortable using BUGS.


```
model lidnorm;
 const
  N = 6;                                        #NUMBER OF STUDIES
 var
   diff[N], vdiff[N], sinv[N],                  #DECLARING VARIABLES
   theta[N], mu, sigma, tau;

 data diff, vdiff in "lidnorm.data";            #READ DATA IN FROM FILE
 inits in "lidnorm.in";                         #STARTING VALUES IN FILE LIDNORM.IN
{
  for (i in 1:N) {
       sinv[i]      <- 1/vdiff[i];             #TRANSFORM TO GET PRECISIONS
       diff[i]       ~ dnorm(theta[i],sinv[i]); #DIFF ~ N(THETA_i,s_i^2)
       theta[i]      ~ dnorm(mu,sigma);         #POPULATION MEAN IS MU AND
                                                #PRECISION SIGMA
} mu         ~ dnorm(0.0,1.0e-6);               #SPECIFY DISTRIBUTION FOR
  sigma      ~ dgamma(0.001,0.001);             #HYPERPARAMETERS
  tau       <- 1/(sigma);
 }
```

Figure 8. BUGS code for the lidocaine meta-analysis. Model statements for estimating a random-effects model in BUGS: mu represents θ, the population mean; sigma is $1/\tau^2$, the precision of the study-specific effects. Comments are proceeded by a # symbol

5.2.1. An example

Figure 8 contains the code for fitting the model

$$d_i \mid \theta_i, s_i^2 \sim \mathrm{N}(\theta_i, s_i^2)$$

$$\theta_i \mid \mu, \tau^2 \sim \mathrm{N}(\mu, \tau^2)$$

$$\mu \sim \mathrm{N}(0, 1e6) \text{ and } \sigma = 1/\tau^2 \sim \mathrm{gamma}(0{\cdot}001, 0{\cdot}001)$$

to the lidocaine data where $\mu = p_{\mathrm{T}} - p_{\mathrm{C}}$, $\theta_i = p_{\mathrm{T}i} - p_{\mathrm{C}i}$, and $d_i = \hat{p}_{\mathrm{T}i} - \hat{p}_{\mathrm{C}i}$. The data are in a matrix format (2 columns, 6 rows) in the file called `lidnorm.data`. Starting values for μ, σ, and $\{\theta_i\}$ were set to 0, 1, and $\{0,0,0,0,0,0\}$, respectively and are in a file called `lidnorm.in`.

The model was run in the background using the `backbugs` command. A portion of the BUGS output based on 2000 iterates after a *burn-in*, defined as an initial run of iterations, of 1000 iterations is displayed in the log-file in Figure 9. The entire log-file is presented in the Appendix. A 95 per cent credible interval for the combined risk difference, μ, is given by $(-0{\cdot}017, 0{\cdot}073)$ with posterior mean estimated as 0·0274. The median of the estimated posterior distribution for the inter-study variation, τ^2, is 0·0011. The study-specific effect corresponding to study 2, θ_2, is estimated as 0·021 with standard deviation 0·034.

6. REPORTING

Once the analytic stages of the meta-analysis have been completed, the results must be reported. A structured abstract summarizing the study objectives, the operational definitions of the treatment


```
Welcome to BUGS on 24 th Sep 1996  at 16:23:8
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.501 for SPARC.
####OUTPUT REMOVED HERE ################
compilation took  00:00:00
Bugs>update(1000)      time for    1000   updates was  00:00:00
Bugs>monitor(mu)
Bugs>monitor(tau)
Bugs>monitor(sigma)
Bugs>monitor(theta)
Bugs>update(2000)      time for    2000   updates was  00:00:01
Bugs>
Bugs>stats(mu)
                  mean        sd       2.5% :  97.5% CI    median      sample
                2.742E-2   2.209E-2  -1.716E-2   7.294E-2   2.691E-2     2000
Bugs>stats(tau)
                  mean        sd       2.5% :  97.5% CI    median      sample
                1.734E-3   2.443E-3   2.745E-4   7.012E-3   1.101E-3     2000
Bugs>stats(theta)
                  mean        sd       2.5% :  97.5% CI    median      sample
[1]             2.710E-2   2.881E-2  -3.069E-2   8.544E-2   2.656E-2     2000
[2]             2.088E-2   3.434E-2  -4.806E-2   8.696E-2   2.143E-2     2000
[3]             2.263E-2   2.314E-2  -2.132E-2   6.747E-2   2.220E-2     2000
[4]             2.246E-2   2.518E-2  -2.727E-2   7.116E-2   2.284E-2     2000
[5]             3.179E-2   2.264E-2  -1.247E-2   7.577E-2   3.165E-2     2000
[6]             3.725E-2   2.072E-2  -4.153E-3   7.889E-2   3.727E-2     2000
Bugs>stats(sigma)
                  mean        sd       2.5% :  97.5% CI    median      sample
                1.172E+3   9.322E+2   1.417E+2   3.626E+3   9.068E+2     2000
Bugs>q()
```

Figure 9. Portion of the BUGS log file for lidocaine example

and population, the meta-analytic design, the search strategy, the results, and the implications for clinical practice as well as for clinical research should be developed. Abstracts that appear in *The Journal of the American Medical Association* are an excellent guide for summarizing the important features of a meta-analysis. Within the body of the text, the data collection measures should be reported in detail as well as the potential biases in the retrieved literature. What was the estimated size of all possible studies to be collected? What were the recall and precision of the retrieved literature? A table of the key elements of each primary study should be included and the overall treatment effect compared to the effects reported in the primary studies qualitatively. A graphical display of the primary data and the estimated pooled effects must be included in the report. The clinical significance of the statistical results should be clearly stated. Do the results imply that lidocaine may increase mortality, for example? Moreover, implications of the results for future research need to be emphasized. The process of data synthesizing permits researchers to identify areas where more research is needed. Are there hospitals for which there was strong evidence in

decrease length of stay even though the overall estimate covers 0? Finally, the methodological limitations should be stated.

7. PROPHYLACTIC LIDOCAINE USE IN HEART ATTACKS

The objective of the meta-analysis is to determine whether there is a detrimental effect of lidocaine on mortality for hospitalized patients with a confirmed heart attack. The primary data include six studies and are reported in Table I. To begin, assume that each estimated risk difference, d_i, is

$$d_i \sim \mathrm{N}(p_\mathrm{T} - p_\mathrm{C}, s_i^2), \quad i = 1, 2, \ldots, 6.$$

7.1. Homogeneity of study means

A test of equality of means yields $Q_W = \sum_i^k W_i(Y_i - \bar{Y}_w)^2 \sim \chi^2_{k-1} = 0{\cdot}86$ and $\chi^2_{0{\cdot}95,5} = 11{\cdot}07$. The null hypothesis of homogeneous study means would *not* be rejected at the 5 per cent level so that the analyst may conclude that the *differences between the studies are so small that any differences are negligible.*

7.2. Fixed-effects results

Table V displays the study-specific risk differences, variances, samples sizes and precisions corresponding to a total of 1106 patients. The study with the largest weight, W_i, corresponds to the study with the largest number of patients (study 6; 300 patients). Using a fixed-effects model, the combined estimate of $p_\mathrm{T} - p_\mathrm{C}$ is $\bar{d}_W = 2{\cdot}94$ per cent with $\mathrm{var}(\bar{d}_W) = 1/5855{\cdot}5 = 0{\cdot}0171$ per cent. A 95 per cent CI for the risk difference is (0·4 per cent, 5·5 per cent), indicating an increase in mortality during the treatment period for lidocaine recipients. Code for calculating the combined estimate and its corresponding standard error using the S-plus software is contained in the Appendix.

7.3. Random-effects results

Two random-effects models were estimated: a model assuming that the distribution of study effects arise from a Normal distribution and a model in which it was assumed that the study effects arise from a Student-t distribution with 4 degrees of freedom. This latter model permitted the tails of the distribution for the random effects to be heavier than the former. For example,

Table V. Lidocaine example. Calculations for the fixed-effects model. $W_i = 1/s_i^2$ is the known weight in the fixed-effects analysis

Study	d_i	$s^2_{d_i}$	n_i	W_i	$\frac{W_i}{\sum W_i}$	$\frac{W_i \times d_i}{\sum W_i}$
1	0·028	0·001778	82	563·1	0·0962	0·002695
2	0·000	0·003757	88	266·2	0·0455	0·000000
3	0·020	0·000813	217	1229·7	0·2100	0·004139
4	0·018	0·001090	203	917·5	0·1567	0·002814
5	0·035	0·000801	216	1248·2	0·2132	0·007532
6	0·044	0·000613	300	1630·7	0·2785	0·01226
Total			1106	5855·5	1·0000	$\bar{d}_W = 0{\cdot}02944$

$E(p_{Ti} - p_{Ci} \mid p_T - p_C, \tau^2, 4) = p_T - p_C$ and $\text{var}(p_{Ti} - p_{Ci} \mid p_T - p_C, \tau^2, 4) = 2\tau^2$, twice that of a Normal distribution. REML estimates using Proc Mixed were obtained for the model that assumed the study effects are from a Normal distribution; Bayesian estimates using BUGS were obtained for both random-effects models. In the Bayesian analyses, non-informative proper priors for the population mean and variance were assumed for both random-effects models. In particular, $p_T - p_C$ was given as $N(0{\cdot}0, 1{\cdot}0\,e^6)$ and $\tau^{-2} \sim \text{Gamma}(0{\cdot}001, 0{\cdot}001)$. See the article by Smith *et al.*[13] regarding specification of clinical priors for the population hyperparameters. A burn-in period of 1000 iterations were used and inference conducted using the subsequent 2000 iterates.

Convergence was achieved after 14 iterations using REML estimation in SAS and resulted in an estimate of 0 for τ^2. Because Q_W is less than the corresponding degrees of freedom, the method of moments (MOM) estimate of inter-study variation, τ^2, is also 0. Consequently, the estimates for the pooled risk difference using a fixed-effects model, and using the MOM and REML estimates random-effects model are identical.

The Bayesian estimate of τ^2 (posterior mean) is the same whether a Student-t distribution or a Normal distribution is assumed for the study-specific risk differences. As a rough guide, one may examine the interval $\hat{p}_T - \hat{p}_C \pm 1{\cdot}96\hat{\tau}$ to determine whether the magnitude of τ^2 is clinically important. If the distribution of the true risk differences is approximately Normal, then one may expect 95 per cent of the studies to have true risk differences in the range $-3{\cdot}1$ per cent to $-2{\cdot}4$ per cent (using the estimates obtain from the Bayesian-t model). Regardless of the model, the pooled estimate of $p_T - p_C$ is about 3 per cent. Moreover, even though the Bayesian intervals for the pooled estimate include 0, the probability that $p_T - p_C > 0$ is 92 per cent. Thus, there is evidence to believe that there is an increase in mortality for lidocaine recipients.

7.4. Diagnostics

The assumption of normality seems reasonable given the observed sample sizes. The combined estimate of the risk difference pooled over the six studies is approximately 3 per cent regardless of whether a fixed-effects model or a random-effects model is employed and lies in the range of observed primary study summaries. Additionally, the results under the random-effects model do not change if we hypothesize that the study-specific effects arise from a Normal distribution or from a Student-t distribution.

The pooled estimate appears insensitive to any one study. Each row in Table VI displays the pooled risk difference when dropping the corresponding study under a fixed-effects model. For example, the combined estimate of $p_T - p_C$ when dropping study 1 and assuming $\tau^2 = 0$ is 3 per cent. The results were similar when performing the sensitivity analysis using a random-effects model.

7.5. Summary

Table VII displays a comparison of the results using fixed-effects and random-effects models; Figure 10 displays the estimated study-specific risk differences and corresponding 95 per cent confidence intervals based on the primary study data and credible intervals based on the posterior means.

The pooled estimate based on the fixed-effects model appears more precise than that based on the Bayesian model because the former assumes inter-study variation is 0. Conversely, the individual study-specific estimates using the Bayesian model are associated with more precision than the raw data because some of the within-study variability is allocated to between-study variability in the

Table VI. Sensitivity analysis. Fixed-effects estimate of $p_T - p_C$ appear to be insensitive to exclusion of studies. The pooled estimate under the fixed-effects model using all the studies is 2·94 per cent

	Lidocaine study		
Study	d_i	$\bar{d}_W$	$\sqrt{\{\text{var}(\bar{d}_W)\}}$
1	0·028	0·030	0·0137
2	0·000	0·031	0·0134
3	0·020	0·032	0·0147
4	0·012	0·032	0·0142
5	0·035	0·028	0·0147
6	0·044	0·023	0·0154

Table VII. Prophylactic lidocaine after MI. Comparison of estimates of the population effect and between-study variance using fixed-effects and random-effects methods

Method	Parameter	Estimate	Variance	95% CI
Fixed-effects model				
MLE	τ^2		Assumed to be 0	
	$p_T - p_C$	2·94%	$(1·31\%)^2$	(0·4%, 5·5%)
	$p_{Ti} - p_{Ci}$	2·94% $\forall i$	$(1·31\%)^2$	(0·4%, 5·5%)
Random-effects model				
DerSimonian & Laird	τ^2	0	–	–
	$p_T - p_C$	2·94%	$(1·31\%)^2$	(0·4%, 5·5%)
REML	τ^2	0	–	–
$p_{Ti} - p_{Ci} \sim N(p_T - p_C, \tau^2)$	$p_T - p_C$	2·94%	$(1·31\%)^2$	(0·4%, 5·5%)
	$p_{Ti} - p_{Ci}$	2·94% $\forall i$	$(1·31\%)^2$	(0·4%, 5·5%)
Bayesian - (Normal)	τ^2	0·17%	$(0·24\%)^2$	(0·03%, 0·70%)
$p_{Ti} - p_{Ci} \sim N(p_T - p_C, \tau^2)$	$p_T - p_C$	2·74%	$(2·21\%)^2$	(−1·72%, 7·29%)
	$p_{Ti} - p_{Ci}$		See Figure 10	
Bayesian - (t)	τ^2	0·18%	$(0·27\%)^2$	(0·03%, 0·71%)
$p_{Ti} - p_{Ci} \sim t(p_T - p_C, \tau^2, 4)$	$p_T - p_C$	2·75%	$(2·20\%)^2$	(−1·71%, 6·87%)
	$p_{Ti} - p_{Ci}$		See Figure 10	

random-effects model. Note that the REML and MOM study-specific estimates are estimated to be constant and identical to the pooled fixed-effects estimate.

In conclusion, the results suggest that lidocaine administered in the hospital phase may increase mortality among patients with a proven heart attack.

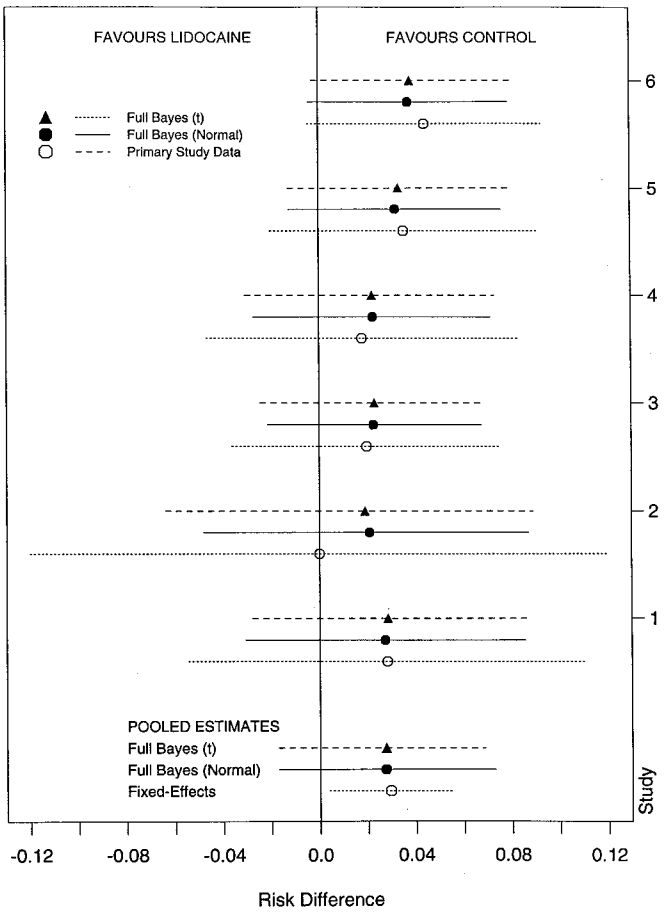

Figure 10. Estimated study-specific and pooled estimates for lidocaine example. Note that $\hat{\tau}^2_{\mathrm{DL}} = \hat{\tau}^2_{\mathrm{R}} = 0$ implying that $(\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}})_{\mathrm{MLE}} = (\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}})_{\mathrm{DL}} = (\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}})_{\mathrm{R}}$. Full Bayes model assumed (I) $p_{\mathrm{T}i} - p_{\mathrm{C}i} \sim \mathrm{N}(p_{\mathrm{T}} - p_{\mathrm{C}}, \tau^2)$ (Normal) or (II) $p_{\mathrm{T}i} - p_{\mathrm{C}i} \sim t(p_{\mathrm{T}} - p_{\mathrm{C}}, \tau^2, 4)$ (Full Bayes (t)). In either case, proper but vague priors were used for the hyperparameters $(p_{\mathrm{T}} - p_{\mathrm{C}}) \sim \mathrm{N}(0, 1{\cdot}0\mathrm{e}^6)$; $\tau^{-2} \sim \mathrm{Gamma}(0{\cdot}001, 0{\cdot}001)$

8. MULTIDISCIPLINARY CARE FOR STROKE INPATIENTS

The objective of this meta-analysis is to determine whether specialist stroke care results in a shorter length of hospitalization compared to routine management on a general medical ward. The data consist of nine randomized trials that are listed in Table II. For each of the $i = 1, 2, \ldots, 9$ studies, we assume that the difference in length of stay is

$$Y_i \sim \mathrm{N}(\mu_{\mathrm{T}} - \mu_{\mathrm{C}}, s_i^2).$$

Table VIII. Stroke example. Calculations for fixed-effects model. $W_i = 1/s_i^2$

Study	Y_i	$s^2_{Y_i}$	n_i	W_i	$\frac{W_i}{\sum W_i}$	$\frac{W_i \times Y_i}{\sum W_i}$
1	−20	40·59	311	0·02464	0·01507	−0·30131
2	−2	2·05	63	0·48856	0·29873	−0·59745
3	−55	15·28	146	0·06544	0·04001	−2·20080
4	−71	150·22	36	0·00666	0·00407	−0·28890
5	−4	20·19	21	0·04952	0·03028	−0·12113
6	1	1·22	109	0·81731	0·49975	0·49975
7	11	95·38	67	0·01048	0·00641	0·07052
8	−10	8·03	293	0·12450	0·07613	−0·76127
9	7	20·69	112	0·04832	0·02955	0·20684
Total			1158	1·63544	1·0000	$\bar{Y}_W = -3·494$

8.1. Homogeneity of study means

A test of equality of study means yields $Q_W = 241·059$. Because $\chi^2_{0·95,8} = 15·5$ the null hypothesis of homogeneous study means would be *rejected* at the 5 per cent level. This is consistent with the message displayed in a plot of the primary study summaries (Figure 2).

8.2. Fixed-effects results

Table VIII displays the study-specific differences in length of hospital stay, variances, samples sizes, and statistics for estimating the fixed-effects pooled estimate. Note that, under a fixed-effects model, the study receiving the largest weight is study 6 having 109 patients (four studies had a larger sample size than study 6). The combined estimate of $\mu_T - \mu_C$ is $\bar{Y}_W = -3·5$ (standard error $= 0·78$) days implying that a decrease in hospital length of stay for stroke patients managed in specialty units compared to stroke patients managed routinely. Code for calculating the combined estimate and its corresponding standard error using the S-plus software is contained in the Appendix.

8.3. Random-effects results

As in the lidocaine example, two random-effects models were estimated: a model assuming that the distribution of differences arise from a Normal distribution and a model in which it was assumed that the differences arise from a Student-t distribution with 4 degrees of freedom. REML estimates using Proc Mixed were obtained for the model assuming the study effects are from the Normal distribution; Bayesian estimates using BUGS were obtained for both random-effects models. In the Bayesian analyses, non-informative proper priors for the population mean and variance were assumed for both random-effects models (for example, $\mu_T - \mu_C \sim N(0·0, 1·0e^6)$ and $\tau^{-2} \sim \text{gamma}(0·001, 0·001)$).

Convergence was achieved after three iterations using REML estimation in SAS and resulted in an estimate of inter-study variation, τ^2, of 685 (standard error $= 380$). Proc Mixed does not provide an interval estimate for τ^2. The MOM estimate of τ^2 is 219, substantially smaller than the REML estimate. Furthermore, the pooled estimate of τ^2 relies on the distributional assumptions and on the mode of inference within the random-effects model. The estimate of $\mu_T - \mu_C$ does vary with the mode of inference but not as much as the estimate of between-study variance. It appears

Table IX. Sensitivity analysis. Posterior summaries of $\mu_T - \mu_C$ and τ^2 when excluding one study at a time. Estimates based on the model that assumes the study-specific effects arise from a Student-t distribution (Bayesian (t))

Study	Y_i	$\hat{\mu}_T - \hat{\mu}_C$	Stroke study Posterior standard deviation	$\hat{\tau}^2$	Posterior standard deviation
1	−20	−9	11	714	1169
2	−2	−13	12	778	1187
3	−55	−5	7	285	554
4	−71	−5	7	259	462
5	−4	−12	12	786	1021
6	1	−13	11	760	1008
7	11	−13	11	693	992
8	−10	−11	12	806	1146
9	7	−14	11	714	1021

Table X. Multidisciplinary care for stroke inpatients. Comparison of estimates for the combined effect and between-study variance using fixed-effects and random-effects models

Method	Parameter	Estimate	Variance	95% CI
Fixed-effects model				
MLE	τ^2		Assumed to be 0	
	$\mu_T - \mu_C$	−3 days	$(1)^2$	(−5, −1)
	$\mu_{Ti} - \mu_{Ci}$	−3 days $\forall i$	$(1)^2$	(−5, −1)
Random-effects model				
DerSimonian and Laird	τ^2	219	–	–
	$\mu_T - \mu_C$	−14 days	$(5)^2$	(−24, −4)
REML	τ^2	685	$(380)^2$	NA
$\mu_{Ti} - \mu_{Ci} \sim N(\mu_T - \mu_C, \tau^2)$	$\mu_T - \mu_C$	−15 days	$(9)^2$	(−32, 2)
	$\mu_{Ti} - \mu_{Ci}$		See Figure 11	
Bayesian (N)	τ^2	892	$(619)^2$	(267, 2591)
$\mu_{Ti} - \mu_{Ci} \sim N(\mu_T - \mu_C, \tau^2)$	$\mu_T - \mu_C$	−15 days	$(10)^2$	(−35, 3)
	$\mu_{Ti} - \mu_{Ci}$		See Figure 11	
Bayesian (t)	τ^2	545	$(478)^2$	(96, 1812)
$\mu_{Ti} - \mu_{Ci} \sim t(\mu_T - \mu_C, \tau^2, 4)$	$\mu_T - \mu_C$	−10 days	$(9)^2$	(−29, 7)
	$\mu_{Ti} - \mu_{Ci}$		See Figure 11	

NA: Confidence intervals are not provided as part of the SAS output

that the hospital stay is between 10 and 15 (± 10)days shorter for those patients managed under specialty care compared to patients managed in the routine manner.

8.4. Diagnostics

Because of lack of homogeneity of means, a fixed-effects model is not appropriate for these data. Each row in Table IX displays the posterior mean and standard deviation of the pooled effect and between-study variance when dropping the corresponding study from the overall analysis. The

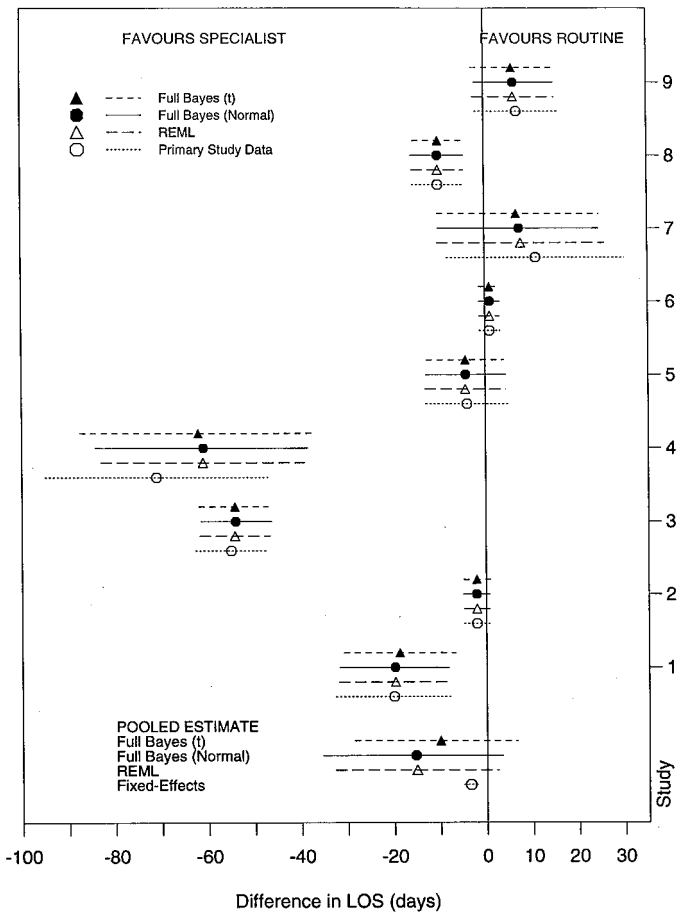

Figure 11. Estimated study-specific and pooled estimates for stroke example. REML estimates converged after three iterations. Full Bayes model assumed (I) $\mu_{Ti} - \mu_{Ci} \sim N(\mu_T - \mu_C, \tau^2)$ (Normal) or (II) $\mu_{Ti} - \mu_{Ci} \sim t(\mu_T - \mu_C, \tau^2, 4)$ (Full Bayes (t)). In either case, proper but vague priors were used for the hyperparameters $(\mu_T - \mu_C) \sim N(0, 1{\cdot}0e^6)$; $\tau^{-2} \sim \text{gamma}(0{\cdot}001, 0{\cdot}001)$)

estimated model employed a Student-t distribution for the underlying study-specific effects and used priors as described earlier. Studies 3 and 4 have an impact on both the estimate of $\mu_T - \mu_C$ and on the estimate of τ^2; when each of these individual studies are excluded from the meta-analysis, the magnitude of the pooled difference is decreased from approximately -11 days to -5 days and the between-study variance is estimated to be about 40 per cent of what it is when all studies are included. This should come as no surprise given the data displayed in Figure 2.

8.5. Summary

Table X displays a comparison of the results using both fixed-effects and random-effects models. Unlike the lidocaine example, the estimates vary widely depending on whether a fixed-effects or a

random-effects model is assumed. However, it is clear that a fixed-effects model is not supported by the data.

Figure 11 displays the estimated study-specific differences in length of stay and corresponding 95 per cent confidence intervals based on the primary study data and credible intervals based on the posterior means. Because the estimate of τ^2 is large compared to the within-study variances, there is little shrinkage of the posterior means to the prior mean (for example, $B_i = 0$) so that the random-effects estimates of $\mu_{Ti} - \mu_{Ci}$ and corresponding credible intervals are almost identical to those inferred by using the primary study data alone.

In conclusion, specialty stroke care generally resulted in shorter lengths of stay compared to routine care. Further investigation is needed to determine which aspects of the specialty care are associated with shorter length of stay and which are generalizable.

9. SUMMARY

In this tutorial, methods were described for combining information across related but independent studies when the primary studies report a single summary statistic such as a mean difference. Analytic methods were confined to those cases in which normality of the summary statistic holds. There are many situations for which combining multivariate statistical summaries will be the primary focus. The methods outlined in this article, however, present an introduction to meta-analysis and a general strategy for inference.

Although three potential sources of variation (within-study, systematic between-study variation, and unexplained between-study variation) were identified as important when synthesizing results, the examples in the tutorial focused primarily on within-study and unexplained between-study variation. It is important to note that systematic between-study variation, as measured through fixed characteristics of the studies, z_i, may exist and can be accounted for in both the fixed-effects and random-effects frameworks. In the case of a fixed-effects model, this would be interpreted to mean that treatment efficacy varies in a systematic fashion between studies characterized by z_i and studies not characterizied by z_i. In the random-effects framework, between-study variation would be modelled by both study characteristics and random variation. This would be interpreted to mean that the efficacy within studies characterized by z_i is more similar than studies without z_i, and moreover, the efficacies within studies characterized by z_i differ because of unexplained between-study variation.

APPENDIX

The S-plus code for performing a fixed-effects analysis is included below for both examples utilized in the tutorial. The complete outputs from two estimation procedures for a random-effects model are also included in this Appendix. The first corresponds to the SAS Procedure Proc Mixed for the stroke study and the second to the BUGS simulation for the lidocaine study.

S-Plus code for the fixed-effects lidocaine meta-analysis

```
gg_matrix(scan("../examples/lidocaine.data"),ncol=4,byrow=T)
deadtreat_gg[,3]; deadcontrol_gg[,4]
ptreat_gg[,3]/gg[,1]; pcontrol_gg[,4]/gg[,2]
ntreat_gg[,1]; ncontrol_gg[,2]
```



```
###COMPUTE RISK DIFFERENCES############

diff_ptreat - pcontrol
vardiff_ptreat*(1-ptreat)/gg[,1] + pcontrol*(1-pcontrol)/gg[,2]
cilower_diff-1.96*sqrt(vardiff); ciupper_diff+1.96*sqrt(vardiff)

## WEIGHTED; FIXED EFFECTS COMBINED ESTIMATE###########

dweights_1/vardiff
dbarweighted_sum(dweights*diff)/sum(dweights)   #combined estimate
vardbarweighted_1/sum(dweights)                 #variance of estimate

print(c(dbarweighted,sqrt(vardbarweighted)))
```

S-Plus code for the fixed-effects stroke meta-analysis

```
los_matrix(scan("../examples/los.data"),ncol=6,byrow=T)
ntreated_los[,1]; meantreated_los[,2]; sdtreated_los[,3]
ncontrol_los[,4]; meancontrol_los[,5]; sdcontrol_los[,6]

difflos_meantreated - meancontrol
pooleds_((ntreated-1)*sdtreated^2 + (ncontrol-1)*sdcontrol^2)/(ntreated +
ncontrol -2)
vardifflos_pooleds*(1/ntreated + 1/ncontrol)
cilower_difflos - 1$\cdot$96*sqrt(vardifflos); ciupper_difflos+1$\cdot$96*sqrt
(vardifflos)

## - WEIGHTED; FIXED EFFECTS COMBINED ESTIMATE###########

losweights_1/vardifflos
losbarweighted_sum(losweights*difflos)/sum(losweights)  #combined estimate
varlosbarweighted_1/sum(losweights)                     #variance of estimate

print(c(losbarweighted,sqrt(varlosbarweighted)))
```

SAS output for the random-effects stroke meta-analysis

```
Stroke Study                              14:31 Tuesday, April 16, 1996   1

                 L       N
                 O   S   C     L
          N      S   D   O     O     S
          T      T   T   N     S     D                     V     S           V
          R      R   R   T     C     C     D               D     T           A
O         E      E   E   R     O     O     I               I     Y   R   C   L
B         A      A   A   O     N     N     F       S       F     I   O   O   U
S         T      T   T   L     T     T     F       P       F     D   W   L   E

1   155   55   47  156    75    64   -20  3155.55    40.586  1   1   1   40.586
2    31   27    7   32    29     4    -2    32.23     2.047  2   2   2    2.047
3    75   64   17   71   119    29   -55   557.33    15.281  3   3   3   15.281
4    18   66   20   18   137    48   -71  1352.00   150.222  4   4   4  150.222
5     8   14    8   13    18    11    -4   100.00    20.192  5   5   5   20.192
6    57   19    7   52    18     4     1    33.27     1.224  6   6   6    1.224
7    34   52   45   33    41    34    11  1597.18    95.376  7   7   7   95.376
8   110   21   16  183    31    27   -10   551.83     8.032  8   8   8    8.032
9    60   30   27   52    23    20     7   576.46    20.694  9   9   9   20.694

Stroke Study                              14:31 Tuesday, April 16, 1996   2
```


The MIXED Procedure

Class Level Information

Class	Levels	Values
STYID	9	1 2 3 4 5 6 7 8 9

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	63.77953195	
1	3	63.33581077	0.00054886
2	1	63.31711993	0.00000851
3	1	63.31684770	0.00000000

Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm	Estimate	Std Error	Z	Pr > |Z|
DIAG	685.09397633	379.91352734	1.80	0.0713
Residual	1.05643873	.	.	.

Model Fitting Information for DIFF

Description	Value
Observations	9.0000
Variance Estimate	1.0564
Standard Deviation Estimate	1.0278
REML Log Likelihood	-39.0099
Akaike's Information Criterion	-40.0099
Schwarz's Bayesian Criterion	-40.0497
-2 REML Log Likelihood	78.0199

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
INTERCEPT	-15.12158889	8.95356233	8	-1.69	0.1297

Solution for Random Effects

Parameter	Estimate	SE Pred	DDF	T	Pr > |T|
STYID 1	-0.27284292	6.21024468	0	-0.04	.
STYID 2	0.03908618	1.42879248	0	0.03	.
STYID 3	-0.87007274	3.87112874	0	-0.22	.
STYID 4	-10.04910369	11.21602341	0	-0.90	.
STYID 5	0.31841048	4.43620482	0	0.07	.
STYID 6	0.02874070	1.10526162	0	0.03	.
STYID 7	3.19213098	9.21507235	0	0.35	.
STYID 8	0.05934993	2.81953230	0	0.02	.
STYID 9	0.64860109	4.48951554	0	0.14	.

Predicted Values

Observed	Predicted	Var Pred	SE Pred	L95M	U95M	Residual
-20.0000	-15.3944	109.7662	10.4769	.	.	-4.6056
-2.0000	-15.0825	81.7301	9.0405	.	.	13.0825
-55.0000	-15.9917	91.6538	9.5736	.	.	-39.0083
-71.0000	-25.1707	177.1315	13.3091	.	.	-45.8293
-4.0000	-14.8032	95.2559	9.7599	.	.	10.8032
1.0000	-15.0928	81.1020	9.0057	.	.	16.0928
11.0000	-11.9295	145.4908	12.0620	.	.	22.9295
-10.0000	-15.0622	86.2581	9.2875	.	.	5.0622
7.0000	-14.4730	95.6211	9.7786	.	.	21.4730

OBS	NTREAT	LOSTREAT	SDTREAT	NCONTROL	LOSCONT	SDCONT	DIFF	SP
1	155	55	47	156	75	64	-20	3155.55
2	31	27	7	32	29	4	-2	32.23
3	75	64	17	71	119	29	-55	557.33
4	18	66	20	18	137	48	-71	1352.00
5	8	14	8	13	18	11	-4	100.00
6	57	19	7	52	18	4	1	33.27
7	34	52	45	33	41	34	11	1597.18
8	110	21	16	183	31	27	-10	551.83
9	60	30	27	52	23	20	7	576.46

OBS	VDIFF	STYID	ROW	COL	VALUE	THETAI	SE_PRED	P_GT_0
1	40.586	1	1	1	40.586	-19.7272	10.4769	0.02986
2	2.047	2	2	2	2.047	-2.0391	9.0405	0.41078
3	15.281	3	3	3	15.281	-54.1299	9.5736	0.00000
4	150.222	4	4	4	150.222	-60.9509	13.3091	0.00000
5	20.192	5	5	5	20.192	-4.3184	9.7599	0.32908
6	1.224	6	6	6	1.224	0.9713	9.0057	0.54294
7	95.376	7	7	7	95.376	7.8079	12.0620	0.74129
8	8.032	8	8	8	8.032	-10.0593	9.2875	0.13938
9	20.694	9	9	9	20.694	6.3514	9.7786	0.74200

BUGS output for the random-effects lidocaine meta-analysis

```
Welcome to BUGS on 24 th Sep 1996  at 16:23:8
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.501 for SPARC.
For general release : please see documentation for disclaimer.
The support of the Economic and Social Research Council (UK)
is gratefully acknowledged.
Bugs>compile("lidnorm.bug")
model lidnorm;
 const
  N = 6;                                          #NUMBER OF STUDIES
 var
   diff[N], vdiff[N], sinv[N],                    #DECLARING VARIABLES
   theta[N], mu, sigma, tau;
 data diff, vdiff in "lidnorm.data";             #READ DATA IN FROM FILE
 inits in "lidnorm.in";                          #STARTING VALUES IN FILE
                                                 LIDNORM.IN
{
  for (i in 1:N) {
       sinv[i]      <- 1/vdiff[i];               #TRANSFORM TO GET PRECISIONS
       diff[i]       ~ dnorm(theta[i],sinv[i]);  #DIFF ~ N(THETA)I,S_I^2)
       theta[i]      ~ dnorm(mu,sigma);          #POPULATION MEAN IS MU AND
                                                 #PRECISION SIGMA
}
  mu         ~ dnorm(0.0,1.0e-6);                #SPECIFY DISTRIBUTION FOR
  sigma      ~ dgamma(0.001,0.001);              #HYPERPARAMETERS
  tau       <- 1/(sigma);
 }
Parsing model declarations.
Loading data value file(s).
Loading initial value file(s).
Parsing model specification.
Checking model graph for directed cycles.
Generating code.
```



```
Generating sampling distributions.
Checking model specification.
Choosing update methods.
compilation took  00:00:00
Bugs>update(1000)      time for    1000   updates was  00:00:00
Bugs>monitor(mu)
Bugs>monitor(tau)
Bugs>monitor(sigma)
Bugs>monitor(theta)
Bugs>update(2000)      time for    2000   updates was  00:00:01
Bugs>
Bugs>stats(mu)
                  mean        sd       2.5% :  97.5%  CI    median    sample
                2.742E-2   2.209E-2  -1.716E-2   7.294E-2   2.691E-2   2000
Bugs>stats(tau)
                  mean        sd       2.5% :  97.5%  CI    median    sample
                1.734E-3   2.443E-3   2.745E-4   7.012E-3   1.101E-3   2000
Bugs>stats(theta)
                  mean        sd       2.5% :  97.5%  CI    median    sample
[1]             2.710E-2   2.881E-2  -3.069E-2   8.544E-2   2.656E-2   2000
[2]             2.088E-2   3.434E-2  -4.806E-2   8.696E-2   2.143E-2   2000
[3]             2.263E-2   2.314E-2  -2.132E-2   6.747E-2   2.220E-2   2000
[4]             2.246E-2   2.518E-2  -2.727E-2   7.116E-2   2.284E-2   2000
[5]             3.179E-2   2.264E-2  -1.247E-2   7.577E-2   3.165E-2   2000
[6]             3.725E-2   2.072E-2  -4.153E-3   7.889E-2   3.727E-2   2000
Bugs>stats(sigma)
                  mean        sd       2.5% :  97.5%  CI    median    sample
                1.172E+3   9.322E+2   1.417E+2   3.626E+3   9.068E+2   2000
Bugs>q()
```

ACKNOWLEDGEMENTS

This work was supported by research grant CA-61141 from the National Cancer Institute. I would like to thank William DuMouchel, AT&T Labs Research, Florham Park, NJ, two anonymous referees and the Associate Editor for helpful comments on earlier versions of the manuscript.

REFERENCES

1. Hine, L. K., Laird, N., Hewitt, P. and Chalmers, T. C. 'Meta-analytic evidence against prophylactic use of lidocaine in Myocardial Infarction', *Archives of Internal Medicine*, **149**, 2694–2698 (1989).
2. Cochran Database of Systematic Reviews, 1995.
3. Cooper, H. M. and Hedges, L. V. (eds). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, 1994.
4. Hedges, L. V. and Olkin, I. *Statistical Methods for Meta-Analysis*, Academic Press, New York, 1985.
5. Rosenthal, R. *Meta-Analytic Procedures for Social Research*, Sage Publications, Beverly Hills, 1984.
6. Cook, D. J., Sackett, D. L. and Spitzer, W. O. 'Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on meta-analysis', *Journal of Clinical Epidemiology*, **48**, 167–171 (1995).
7. Bishop, Y. V., Fienberg, S. E. and Holland, P. W. *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA, 1975, Chapter 6.
8. Sands, M. L. and Murphy, J. R. 'Use of kappa statistic in determining validity of quality filtering for meta-analysis: a case study of the health effects of electromagnetic radiation', *Journal of Clinical Epidemiology*, **49**, 1045–1051 (1996).
9. Cook, T. D. and Campbell, D. T. *Quasi-Experimentation*: *Design & Analysis Issues for Field Settings*, Houghton Mifflin Company, Boston, MA, 1979.

10. Chalmers, T. C, Smith, H. Jr., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. 'A method for assessing the quality of a randomized control trial', *Controlled Clinical Trials*, **2**, 31–49 (1981).
11. Patterson, H. D. and Thompson, R. 'Recovery of interblock information when block sizes are unequal', *Biometrika*, **58**, 545–554 (1971).
12. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
13. Smith, T. C., Spiegelhalter, D. J. and Thomas, A. 'Bayesian approaches to random-effects meta-analysis: a comparative study', *Statistics in Medicine*, **14**, 2685–2699 (1995).
14. Morris, C. N. and Normand, S. L. 'Hierarchical models for combining information and for meta-analyses', *Bayesian Statistics 4*, Oxford University Press, 1992, pp. 321–344 (with Discussion).
15. DuMouchel, W.H. 'Bayesian meta-analysis' in *Statistical Methodology in the Pharmaceutical Sciences*, Marcel Dekker, New York, NY, 1990, pp. 509–529.
16. DuMouchel, W.H. and Harris, J.E. 'Bayes methods for combining the results of cancer studies in humans and other species', *Journal of the American Statistical Association*, **78**, 293–315 (1983).
17. Morris, C.N. 'Approximating posterior distributions and posterior moments', *Bayesian Statistics 3*, Oxford University Press, 1988, pp. 327–344 (with Discussion).
18. DerSimonian R. and Laird, N. 'Meta-Analysis in Clinical Trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
19. Light, R. J. and Pillemer, D. B. *Summing Up*: *The Science of Reviewing Research*, Harvard University Press, Boston, MA, 1984.
20. Iyengar, S. and Greenhouse, J. 'Selection models and the file-drawer problem' with Discussion, *Statistical Science*, **3**, 109–135 (1988).
21. Dear, K. B. and Begg, C. B. 'An approach for assessing publication bias prior to performing a meta-analysis', *Statistical Science*, **7**, 237–245 (1992).
22. Hedges, L. V. 'Modeling publication selection effects in meta-analysis', *Statistical Science*, **7**, 246–255 (1992).
23. Givens, G. F., Smith, D. D. and Tweedie, R. L. 'Estimating and adjusting for publication bias using data augmentation in Bayesian meta-analysis', Technical Report, Department of Statistics, Colorado State University, 1995.
24. Statistical Sciences, *S-PLUS User's Manual*, Version 3.1 Supplement, Statistical Sciences, Inc., Seattle, WA, 1992.
25. Normand, S. L. 'Meta-analysis software: a comparative review', *American Statistician*, **49**, 298–309 (1995).
26. DuMouchel, W. H. 'Predictive cross-validation of Bayesian meta-analysis', in *Bayesian Statistics 5*, Oxford University Press, 1996, pp. 107–128, (with Discussion).
27. Spiegelhalter, D. J., Thomas, A. and Best, N. G. 'Computation on bayesian graphical Models', in *Bayesian Statistics 5*, Oxford University Press, 1996, pp. 407–425 (with Discussion).
28. Normand, S. L. Discussion of 'Computation on Bayesian Graphical Models' by Spiegelhalter, D. J., Thomas, A. and Best, N. G. in *Bayesian Statistics 5*, Oxford University Press, 1996, pp. 420–421.

