NISS

Hierarchical Bayes Linear Models for Meta-Analysis

William DuMouchel

Technical Report Number 27
September, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Research supported by the U. S. Environmental Protection Agency Cooperative Agreement CR# 819638-01-0 with the National Institute of Statistical Sciences.

Hierarchical Bayes Linear Models for Meta-Analysis

William DuMouchel

Columbia University
Division of Biostatistics
dumouch@bayes.cpmc.columbia.edu

September 1994

Abstract

This paper develops and illustrates the use of hierarchical Bayes linear models for meta-analyses. The methodology can be thought of as a compromise between the fixed-effect meta-analytic methods that are most often used in the literature (properly criticized for ignoring sources of among-study variation) and the more extreme critics of meta-analysis who argue that it is almost never appropriate to combine results from disparate studies. After reviewing the recommendations of a recent NAS report on combining information, the paper explains the data requirements, the statistical models and the prior distributions used in the hierarchical Bayes approach. Much of the focus is on the estimation and interpretation of the standard deviation, τ, of interstudy differences in effects. To this end, a special plot, called a *trace plot*, is introduced that shows the role of τ in the meta-analysis. Another useful graph summarizes the "shrinkage" property of the Bayesian posterior distributions of the study-specific effect estimates. The methodology is illustrated with a meta-analysis of 9 studies of the effect of indoor air pollution on childhood respiratory illness. Two analyses of these data (originally collected and analyzed by Hasselblad et al., 1992) are presented. The first ignores differences in study designs, while the second uses characteristics of the studies (namely, the list of potential confounders each study considered) to build a model to explain differences in effects among studies, and to estimate the average effect that would be found in studies that adjust for all three of the potential confounders being considered.

1. Introduction

1.1 The problem

Traditionally, statistical theory and methods have focused on the analysis of data from a single source. An experiment or a survey is designed and carried out by a single person or team, and the resulting data are analyzed within a statistical model not designed to encompass other studies. This focus is natural when the study being analyzed is unique or is intended to be definitive – to supersede previous studies of the same problem, and to be independent of studies of related problems. For example, previous studies may have been subject to potentially grave biases from flaws in study design or execution. The natural desire has been to avoid contamination of results of a current analysis by including possibly irrelevant or incorrect information. Another possible motivation for ignoring other data in the analysis is a desire to present an independent replication of findings from earlier studies.

Increasingly, this restricted view is recognized as inadequate. The scope and quantity of research in all fields of science are so great these days that every important problem is being attacked by multiple researchers who want and need to build on each other's results and on the results of previous work. There may be tens or hundreds of studies whose results are relevant to a particular problem, such as the evaluation of a new surgical procedure or the assessment of an environmental risk. Although no study can be perfect, it doesn't make sense to completely ignore the pattern of results from previous studies. Two approaches to the use of information from previous studies are the subjective narrative literature review, on the one hand, and, at the other extreme, the simple pooling of data from different studies. Sole reliance on either extreme is poor science and wasteful of resources.

The narrative review approach shies away from a quantitative analysis of data from multiple studies in favor of a discussion of the literature to put the different studies "in context." This discussion often dismisses the majority of previous studies as being either too flawed or too distant from the current problem under consideration to be useful. The remaining few studies are then brought up one by one and evaluated as either supporting or failing to support the author's own study or point of view. For studies in the latter category, possible reasons for the different results, such as differences in populations or treatments or conditions, are hypothesized. The reviewer usually recommends further studies to confirm or disprove such hypotheses. This approach is liable to be quite subjective, leading to different conclusions when different reviewers discuss the same body of studies. The neglect of formal statistical tools like numerical summaries or uncertainty measures integrating the study results is a step backwards from the preferred way of drawing conclusions from data. If confidence intervals, for example, are restricted to relying on the results from a single study, how can we express numerically the increased confidence we feel when results are replicated across studies?

At the opposite extreme from avoiding all quantitative syntheses of statistical results from different studies is the simple pooling of data from them. For example, a regression or an analysis of variance may be performed, and treatment comparisons made as if only one study were involved. In situations where the raw data is not available from the individual studies, it may be possible to combine summary statistics from different studies in a weighted analysis that duplicates the result of the treatment comparison that would have resulted from pooling the raw data. This is what often results from a fixed-effect meta-analysis. The word "meta-analysis" denotes a quantitative analysis combining the results from different studies, and the adjective "fixed-effect" refers to the implicit assumption that each study estimates exactly the same treatment effects without bias. In fact it is not likely that such a naive assumption can be justified, given that the studies have been designed and carried out by different investigators, with different populations and under different conditions, at different times, and perhaps even with different purposes in mind. A more sophisticated approach to producing an integrated analysis is required.

A situation that is conceptually similar to the combining of information from different studies is the analysis of data from a multi-site study. For example, a large clinical trial may be carried out by many cooperating teams of researchers, dispersed around a country or even several countries. The purpose of the cooperation is to make the results from each site as comparable as possible, to ensure validity of across-site summaries and between-site contrasts. But, inevitably, results from different sites will lose some degree of comparability over the course of a long study, as each team of researchers copes with the site-specific difficulties that tend to arise. Again, the overall statistical analysis must be more ambitious than a recital of the estimation results at each site, and more sophisticated than a pooled analysis that assumes that the only source of random error is sampling variation within each site.

This chapter is primarily concerned with methods for combining information from different studies, given that they are at hand. There is little or no discussion of how to find or select studies to include in the analyses. Often the studies being integrated are only a fraction of all the studies that are potentially relevant to the scientific issue being investigated. One must then distinguish between producing an adequate summary of the studies at hand and claiming that the resulting inferences are a definitive summary of all research on the problem. Of particular concern is the so-called "file-drawer problem" (Rosenthal 1979), which occurs if those studies which happen to show statistically significant effects have a much greater chance of being published than do studies without significant effects, which remain hidden in their investigators' file drawers. Among many other requirements, to be definitive, a review of the research literature must be conducted like a formal statistical survey: (1) define a protocol for inclusion into the review, (2) document the search for studies and list those found and those excluded for various reasons, (3) address questions of potential combinability and biases of the studies used, and (4) define a recognized statistical methodology to be used to combine the results. This chapter is primarily about steps (3) and (4). See Light and Pillemer (1984) or Chalmers et al (1987ab) for more information about steps (1) and (2).

1.2 Statistical Models and Methods

Before introducing the hierarchical Bayes linear model for combining information, a brief glance at other proposed methods is in order. The methods of descriptive statistics, such as summary tables and graphs that allow an overview and comparison of results from the different sources, are very important and necessary to give the meta-analyst a "feel" for the analysis. In some cases, there is no need to go on to a more formal meta-analysis. Usually, however, the analysis continues with an inferential phase involving the formulation and fitting of a statistical model.

There are both classical and Bayesian inferential methods of meta-analysis. The book by Hedges and Olkin (1985) surveys classical statistical methods of combining information. The articles by Louis, Fineberg and Mosteller (1985) and Sacks et al. (1987) each review many meta-analyses in the areas of public health and clinical trials. Many classical statistical methods for combining studies have been used. They include merely counting the number of studies which find a significant result, computing an average effect size across all studies, and computing a single combined level of significance for a set of statistical hypothesis tests, one from each study. Unfortunately, methods for combining significance levels from different studies do not distinguish between small studies with large effects and large studies with small effects. In general, the well-known disadvantages inherent in the excessive use of hypothesis testing during data analyses come home to roost in attempts to combine them. For example, two studies may yield roughly similar treatment effects, but if one study shows a significant treatment difference ($p < .05$) while the other study shows no significant difference ($p > .05$), a research synthesis may try to "explain" the results by focusing on characteristics in which the studies differ.

According to Gaver et al (1992, p. 1, emphasis in original)

- **Methods for combining information based only on P-values from each information source should be discontinued**, in favor of estimates of quantities of direct scientific or decision-making relevance (such as the effect of a drug on mortality), together with uncertainty assessments for those estimates.

Classical methods for combining information along the lines of this suggestion amount to taking a weighted average of the estimates from each study, where the weights depend variously on the uncertainty assessments attached to each estimate.

Sometimes, especially in the social sciences, the studies do not yield hard estimates of easily interpreted quantities having easily understood units, but instead treatment effects are represented by correlation coefficients or a standardized score whose interpretation is more ambiguous across studies or outside an experimental or survey context. Meta-analyses based on such quantities are certainly to be preferred to those based only on P-values, but are not likely to be as satisfying as those based on estimates of quantities of direct scientific or decision-making relevance.

The Bayesian approach is usually distinguished by having less emphasis on hypothesis testing, as opposed to estimation. As discussed in more detail below, the hierarchical Bayesian approach is distinguished by the construction and use of a formal statistical model at two levels. At the first level, a parametric model is set up for each of the individual studies, in which a likelihood function relates the distribution of the sample statistics to one or more unknown parameters characterizing that study. At the second level, a parametric statistical model is constructed to relate the parameters from the separate studies to each other. This makes sense in a Bayesian formulation, since in the Bayesian approach parameters are not just fixed unknown constants; they are also thought of as random variables having distributions of their own. The computations involved in combining the two levels of the model are derived by the application of Bayes' formula for conditional probability. The Bayesian analysis provides a posterior distribution for each of the study-specific parameters; this will be a satisfying goal if the models are defined in terms of meaningful parameters. Research questions must be phrased as questions about the values of these parameters. For further discussion of the Bayesian approach to meta-analysis, see Morris and Normand (1992).

Fixed versus random effects

An important distinction is whether the method of synthesis assumes that each study is measuring the same underlying parameter. For example, when combining the results from several studies of an occupational risk, are the "true" dose-response curves assumed to be the same across studies? A common recommendation is that one should first conduct a formal hypothesis test of between-study homogeneity; only if this test is not significant is one advised to estimate the supposedly common parameter uniting the studies. This recommendation ignores the distinction between failing to reject interstudy homogeneity and proving that it is present. Methods of synthesis which explicitly allow for a component of variance for study-to-study variation avoid this logical fallacy. Fitting such more complex models sometimes requires more detailed information from the primary studies than do other combining information techniques, but they also usually provide more useful results for scientific and policy purposes. All Bayesian meta-analytic models involve random effects, but the converse is not true, since empirical Bayes models and other classical techniques can include random effects in their formulation.

A fixed-effect model can allow different parameters for different studies by assuming that fixed characteristics of the studies determine the parameter values. For example, the true treatment effect for studies of males may differ from that of studies of females, but a fixed-effect formulation still assumes that two very large studies on the same gender will agree up to within-study sampling error. It is also possible to have models for combining information that include both fixed between-study differences and random between-study differences—these are called *mixed models.* The hierarchical Bayesian model recommended here allows for both types of between-study variation. Section 2 considers the simpler situation where there are no study-level covariates being considered, while Section 3 considers the more general situation.

The National Academy of Sciences report, *Combining Information: Statistical Issues and Opportunities for Research*, (Gaver et al. 1992), referenced above for its recommendation to avoid meta-analyses based on P-values, also provides several other recommendations, four of which are (p. 2, emphasis in original):

- Current CI [Combining Information] practices would also be improved if researchers were more explicit in **model** formulation concerning how the model expresses judgments on the **similarity (exchangeability)** and on the differences of information sources (subjects, variables, research studies, bodies of expert opinion) to be combined.

- **Hierarchical statistical models** are a useful framework for CI. Their use in fields in which they are not yet routinely employed is to be encouraged, as is an increase in the coverage of such models in intermediate and advanced statistics courses.

- CI modeling would be improved by an increase in the use of **random effects** models in preference to the current default of **fixed effects** models. At a minimum, the panel believes that researchers will often find it useful to perform a sensitivity analysis in which both kinds of models are fit and the substantive conclusions from the two approaches are compared.

- A general-purpose statistical computing package allowing investigators to routinely perform **interactive Bayesian analyses** in hierarchical models would gain immediate and widespread acceptance and would help promote the use of such models.

The development of this paper follows these recommendations.

2. Exchangeable Sets of Studies

2.1 Data Requirements

Suppose that the results from K ($K > 1$) studies are available, and it is wished to combine these results in an analysis that summarizes and accounts for differences between them. First we discuss the data requirements of the hierarchical Bayesian approach and then describe the model.

Estimates of parallel quantities from each study

To carry out a hierarchical Bayesian analysis, the data available from each must have a certain commonality or comparability. For example a series of carcinogenesis studies may each estimate a relative risk or a standardized mortality ratio.

In the social and behavioral sciences, the term "effect size" often has a specific meaning related to a convenient way of relating the power of an hypothesis test to the sample size. See, for example, Cohen (1977). In a comparison of means for treatment versus control groups, the effect size might be defined as the difference in means divided by the standard deviation of the response within the control population. One question is whether different studies are more usefully compared by their effect sizes thus defined, or by the simple difference in means unscaled by the response standard deviation. For example, in a review of studies of treatments for depression, the treatment effect may be measured as the improvement in the patients' average Hamilton score, a widely-used scale for depression that is based on a patient questionnaire. Assuming that every study reported Hamilton score mean changes, the raw Hamilton changes provide a metric that is relatively well understood in the psychiatric research community. On the other hand, if different studies are based on different instruments measuring the extent of depression, the scaling of each improvement score by a study-specific, or at least an instrument-specific, standard deviation is advisable.

The estimate taken from the i^{th} study will be denoted y_i, $i = 1, \ldots, K$. It is assumed that the expectation of y_i is a study-specific parameter θ_i. The column vectors of the y_is and of the θ_is are denoted $\boldsymbol{y}$ and $\boldsymbol{\theta}$, respectively.

Standard Error for Each Study's Estimate

The second requirement which the data must satisfy is that each of the estimates must be accompanied by its (approximate) standard error, which will be denoted by s_i, $i = 1, \ldots, K$.

The formal Bayesian model assumes that the estimates from each study have normal distributions with known variances, conditional on the true parameter value. That is, $y_i \mid \theta_i \sim N(\theta_i, s_i^2)$. If the individual estimates are not based on very small samples, this assumption is likely to be a good one, and in any case moderate violations of this assumption will not seriously affect the analysis.

A more serious consideration is the fact that many published studies do not contain enough statistical detail to enable the parameter estimates and, especially, their standard errors, to be calculated. A particular author may have taken a different analysis tack than that required by the proposed meta-analysis, or even misanalyzed the data. To construct the needed set of parallel parameter estimates with associated standard errors it may be necessary to reanalyze many of the original data sets. This would require the cooperation of the original researchers and might involve more time and effort than would the combining information step. In addition, the construction of the Bayesian model can be constrained by the types of data available from the studies being reviewed. Of course, in many combining information contexts, such as the analysis of data from a multi-site clinical trial, access to the original data from each site, and close conformance of data definitions across sites, is taken for granted.

2.2 Hierarchical Model

This Section concerns the combining of information when there are no covariates to distinguish the studies, although there may be greater variation in the y_i than can be explained by sampling error. That is, the model for y_i, $i = 1, ..., K$, is

$$y_i = \mu + \delta_i + \varepsilon_i \tag{1a}$$

$$\theta_i = \mu + \delta_i \tag{1b}$$

$$\delta_i \sim N(0, \tau^2) \tag{1c}$$

$$\varepsilon_i \sim N(0, s_i^2) \tag{1d}$$

where the δ_i and the ε_i are assumed to be independent of each other and across studies. Viewed as a classical variance-component model, there are two parameters, μ and τ, to estimate. In addition, the θ_i can also be estimated, if study-specific estimates are desired.

As mentioned previously, many classical meta-analyses have omitted the random effect δ_i, or, equivalently, have assumed that $\tau = 0$. The question of whether to include a random effect has generated controversy (Peto, 1987). Some analysts recommend always choosing $\tau = 0$. The most common non Bayesian recommendation, see, e.g., DerSimonian and Laird (1986), advocates formally testing the hypothesis $\tau = 0$. If the hypothesis is rejected, the test statistic can be used to estimate τ, and then the analysis proceeds as if τ were known. This recommendation is certainly better than assuming $\tau = 0$ regardless of the data. But there are two major weaknesses in this approach. First, as mentioned above, it is not correct to assume $\tau = 0$ just because the hypothesis cannot be rejected, especially when K, the number of studies being combined, is not large ($K < 20$, say) since the power of the test is then small. Second, even when the hypothesis is rejected and τ must be estimated, the uncertainty in τ is not taken

into account during the subsequent analysis. This can lead to unduly optimistic estimates of the uncertainty in μ and θ.

The Bayesian analysis described next overcomes these difficulties by assuming a prior distribution for τ and then using its posterior distribution to properly account for uncertainty in τ. In order to do this, it helps to be able to compute and display the posterior distribution of τ and the dependence of μ and θ on τ.

The Bayesian specification is completed by adding the prior distributions for μ and τ:

$$\mu \sim N(m, d^2 \to \infty) \tag{2a}$$

$$\tau \sim \pi(\tau) \tag{2b}$$

In the absence of more specific information, we use the *diffuse prior* distribution for μ, defined as (2a) with arbitrary m and arbitrarily large variance, namely $d = \infty$. This *noninformative prior* distribution is a common Bayesian device for estimating a location parameter. For the unknown standard deviation, we will use a *proper prior* distribution for τ, namely, a $\pi(\tau)$ that satisfies the requirements of a true probability distribution: $\pi(\tau) \geq 0$, $\int \pi(\tau)\, d\tau = 1$. For now, a log-logistic prior distribution for τ will be assumed, namely

$$\pi(\tau) = s_0 / (s_0 + \tau)^2 \tag{3}$$

where

$$s_0^2 = K / \Sigma s_i^{-2} \tag{4}$$

Section 2.4 discusses this and other choices for $\pi(\tau)$. The density (2.3) has median equal to s_0, where s_0^2 is the harmonic mean of the K sampling variances s_i^2, and is extremely highly dispersed, since the expectation of both τ and $1/\tau$ are infinite.

Posterior Distributions

The Bayesian analysis proceeds by computing the posterior distributions of μ, τ, and the θ_i given the data $y_1, \ldots, y_K$. The formulas behind these computations are given in the Appendix. In some situations, primary interest centers on estimation of the average study effect μ, while the individual study effects θ_i are of lesser interest. In the other situations, the reverse may be true. Denote the posterior distribution of τ by $\pi(\tau \mid y)$. Graphs of $\pi(\tau \mid y)$ display the dependence of the meta-analysis on this crucial variance component. Small values of τ describe situations in which the meta-analysis is powerful and studies can "borrow strength" from each other. Very large values of τ imply that not much can be gained from

combining the studies. The values of $\pi(\tau \mid y)$ tell us how to weight the results, corresponding to different values of τ, in the final analysis.

2.3 Example: Nine Nitrogen Dioxide Epidemiology Studies

Exhibit 1 shows a graph of 95% confidence intervals from nine studies of the effects of nitrogen dioxide on the odds of respiratory illness in children, taken from Hasselblad et al (1992) and Kotchmar (1993). Although the designs of the studies differed in several important dimensions, consideration of these differences will be deferred until Section 3.

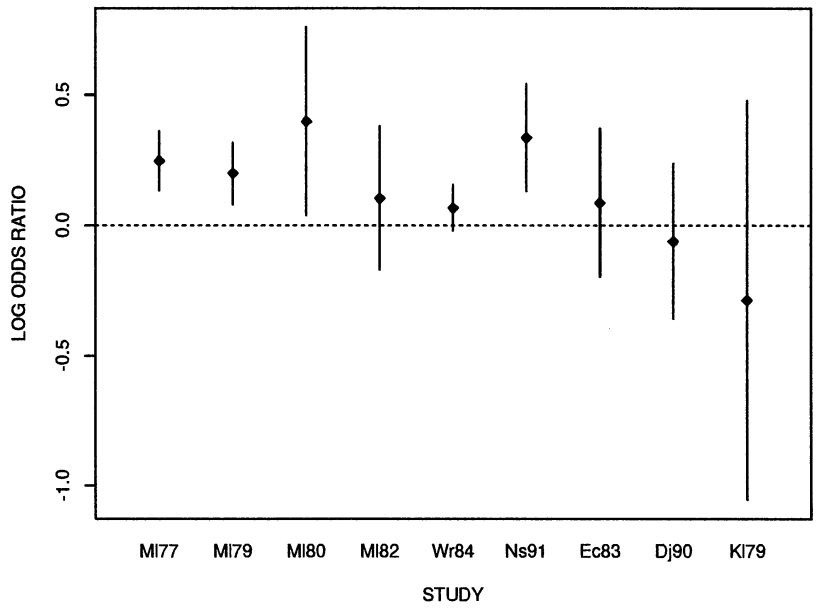

Exhibit 1. 95% confidence intervals from 9 studies of the effects of NO_2 on childhood respiratory illness

For each study, we define

$$y_i = \log_e (OR_i)$$

$$s_i = \log_e (Upper_i/Lower_i) / 3.92$$

where OR_i is the reported odds ratio associating indoor NO_2 exposure (the presence of a gas stove) to childhood respiratory illness in study i, $i = 1, ..., 9$, and $(Lower_i, Upper_i)$ is the reported 95% confidence interval for the true odds ratio in study i, denoted by θ_i. The intervals graphed in Exhibit 1 are $y_i \pm 1.96s_i$. Note that some confidence intervals are much longer than others, because of the widely differing study sizes. Five of the nine confidence intervals (including that from the largest study) cover $\theta = 0$, implying that those studies failed to show a significant effect. Questions that a meta-analysis should answer are: Do these studies agree within the limits of their sampling errors? What confidence limits can be placed

on μ, the average log odds ratio for all similar future studies? Can the confidence limits on an odds ratio from one of these studies be improved in light of the results from the other studies?

First, the posterior distribution of τ and its interpretation are presented, as well as how estimates of μ and of the θ_i depend on τ. The marginal posterior distributions of μ and of the θ_i are then formed by averaging with respect to the distribution of τ.

Conditioning on τ

From the model (1), if τ is known the variables y_i are independently normally distributed with mean μ and variance $\tau^2 + s_i^2$. If μ has a flat prior distribution, then the posterior expectation of μ, conditional on τ and on $\boldsymbol{y} = (y_1, \ldots, y_K)^t$, is a weighted average of the y_i, where the weights are inversely proportional to $\tau^2 + s_i^2$. That is,

$$\mu^*(\tau) \equiv E[\mu \mid \boldsymbol{y}, \tau] = \Sigma_i \, w_i(\tau) \, y_i \tag{5}$$

where

$$w_i(\tau) = (\tau^2 + s_i^2)^{-1} / \Sigma_j \, (\tau^2 + s_j^2)^{-1} \tag{6}$$

Also, the expectation of each θ_i, given $\boldsymbol{y}$ and τ, is the so-called shrinkage estimate, an average of $\mu^*(\tau)$ and y_i:

$$\theta_i^*(\tau) \equiv E[\theta_i \mid \boldsymbol{y}, \tau] = \mu^*(\tau) + [y_i - \mu^*(\tau)] \, \tau^2/(\tau^2 + s_i^2) \tag{7}$$

Finally, by combining the assumption $y_i \sim N(\mu, \tau^2 + s_i^2)$ with the prior distribution $\pi(\tau)$ for τ, the posterior distribution for τ, $\pi(\tau \mid \boldsymbol{y})$, is computed as in the Appendix. Exhibit 2 shows a histogram representing $\pi(\tau \mid \boldsymbol{y})$, overlaid with a curve representing $\mu^*(\tau)$. The histogram shows that plausible values for τ range from 0.01 up to 0.25 or so. The mean of the distribution is $\tau^* = 0.75$. For comparison, the values of s_i ranged from 0.05 (study Wr84) to 0.39 (study Kl79). As Exhibit 2 shows, the weighted average $\mu^*(\tau)$ stays within the narrow range (0.150, 0.165) for $\tau < 0.25$. Although τ as large as 1 is virtually ruled out by $\pi(\tau \mid \boldsymbol{y})$, the curve in Exhibit 2 shows that assuming such a large value would give more weight to the least precise Kl79 study and lower the estimate of μ.

The marginal posterior mean of μ, denoted μ^*, is computed by multiplying the curve height by the histogram height and summing:

$$\mu^* \equiv E[\mu \mid \boldsymbol{y}] = \int \mu^*(\tau) \, \pi(\tau \mid \boldsymbol{y}) \, d\tau \tag{8}$$

The marginal posterior variance of μ, denoted μ^{**}, is computed by similarly averaging with respect to $\pi(\tau \mid \boldsymbol{y})$, but there are two components being averaged, namely the variance of μ given $\boldsymbol{y}$ and τ, and the square of the deviation of $\mu^*(\tau)$ from μ^*:

$$\mu^{**} \equiv V[\mu \mid y] = \int \{V[\mu \mid y, \tau] + [\mu^*(\tau) - \mu^*]^2\}\ \pi(\tau \mid y)\ d\tau \qquad (9)$$

Finally, the posterior probability $P(\mu > 0 \mid y)$ is also computed by averaging $P(\mu > 0 \mid y, \tau)$:

$$P(\mu > 0 \mid y) = \int \Phi(\mu^*(\tau)/V[\mu \mid y, \tau]^{1/2})\ \pi(\tau \mid y)\ d\tau \qquad (10)$$

where $\Phi(z)$ is the cumulative standard normal distribution function. The formula for $V[\mu \mid y, \tau]$ is found in the Appendix. In the present example, we have $\mu^* \pm \mu^{**1/2} = 0.163 \pm 0.046$, and $P(\mu > 0 \mid y) = 0.998$, so there is very strong evidence that the average of all similar studies will be positive, with a predicted odds ratio of $e^{0.163} = 1.18$.

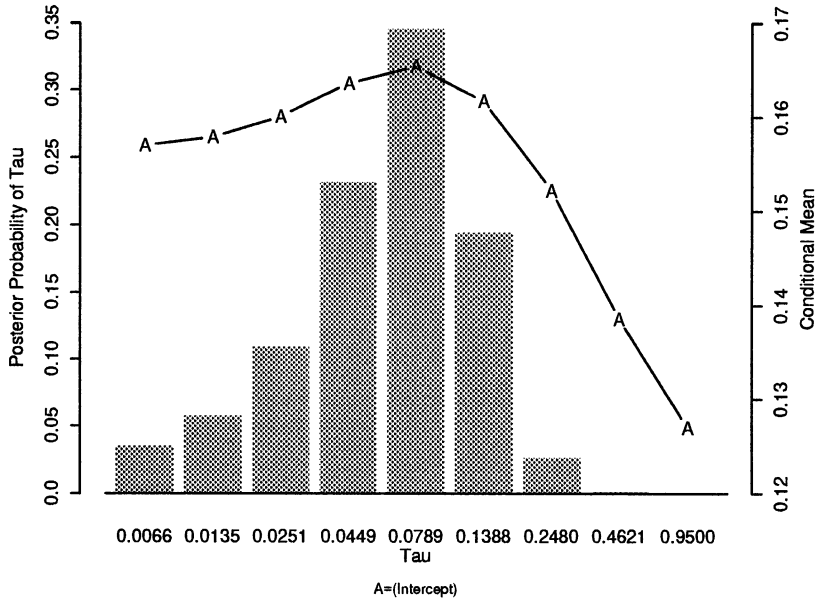

Exhibit 2. Trace plot showing the posterior distribution of τ (histogram and left axis) and the posterior expectation of μ given τ (curve and right axis) for the nine NO_2 studies

Exhibit 3 shows another version of the trace plot for this analysis in which curves representing the individual study estimates $\theta_i^*(\tau)$ for $i = 1, ..., 9$, are also plotted. These curves show why $\theta_i^*(\tau)$ are called shrinkage estimates, since at the right of the plot, when τ is large, the heights of the curves are near the original values of the y_i, while as τ decreases the spread among the curves "shrinks" toward a common point, namely $\mu^*(0)$, as τ approaches 0. (The curve labeled A is $\mu^*(\tau)$.) Note that the rate of shrinkage varies across studies, and studies with larger s_i shrink faster. Therefore the Kl79 study, which had the smallest $y = -0.29$ but the largest $s = 0.39$, is actually estimated to have a larger θ than two other studies, assuming that $\tau < 0.15$ or so.

The marginal estimates of each study's true mean are computed by averaging their shrinkage estimates (7) with respect to $\pi(\tau \mid y)$:

$$\theta_i^* \equiv E[\theta_i \mid y] = \int \{\mu^*(\tau) + [y_i - \mu^*(\tau)]\, \tau^2/(\tau^2 + s_i^2)\}\, \pi(\tau \mid y)\, d\tau \qquad (11)$$

Posterior variances, θ_i^{**}, and probabilities $P(\theta_i > 0)$ for the θ_i are computed using formulas analogous to (9) and (10). For example, the posterior mean and standard deviation for the last, most negative study are 0.142 ± 0.098, with $P(\theta_{Kl79} > 0) = 0.929$, so this Bayesian analysis estimates that a repeat of this study with a larger sample would show a quite positive effect, because of the phenomenon statisticians call "regression to the mean."

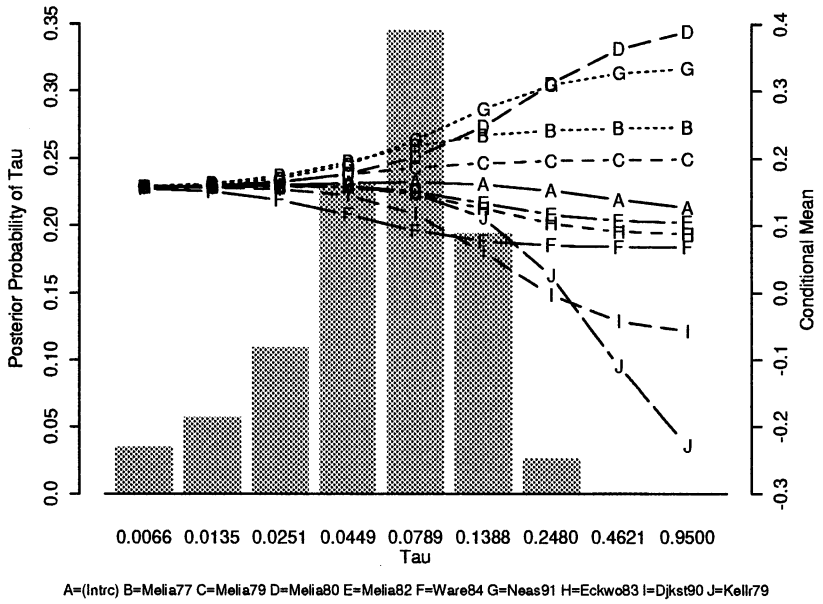

Exhibit 3. Trace plot showing the posterior distribution of τ and the posterior expectations of μ and of θ_i given τ, for each of the nine NO_2 studies

Note that in as much as the values of $\mu^*(\tau)$ and $\theta_i^*(\tau)$ change within the central part of the distribution of τ, the uncertainty in τ contributes to uncertainty in the values of μ and θ_i. This is an important property of the Bayesian method not shared by the classical random-effects meta-analysis. In the latter method, when the number of studies is small, the point estimate of τ can be 0 although large values of τ also agree with the data. In such cases the classical random-effect results reduce to those of the fixed-effect model -- a very over-optimistic assessment of estimation accuracy.

Summary Graph of Hierarchical Bayes Analysis

Exhibit 4 shows a summary graph comparing the Bayesian results to stand-alone estimation. Like the trace plot, this graph shows the pattern of the shrinkage estimation, but unlike the trace plot the focus is on the marginal distribution of each θ_i, after averaging over the distribution of τ. The vertical axis of the plot lists each of the studies being combined, with the addition of an area at the top of the plot, labeled "Prior +/- Tau". On the horizontal axis is a scale for μ, θ and y. At the top is a plotted "X" with a

horizontal error bar displaying the values of $\mu^* \pm \tau^*$. Since the model assumes that each $\theta_i \sim N(\mu, \tau^2)$, this interval contains approximately 68% of the prior probability for each θ_i, given $\mu = \mu^*$.

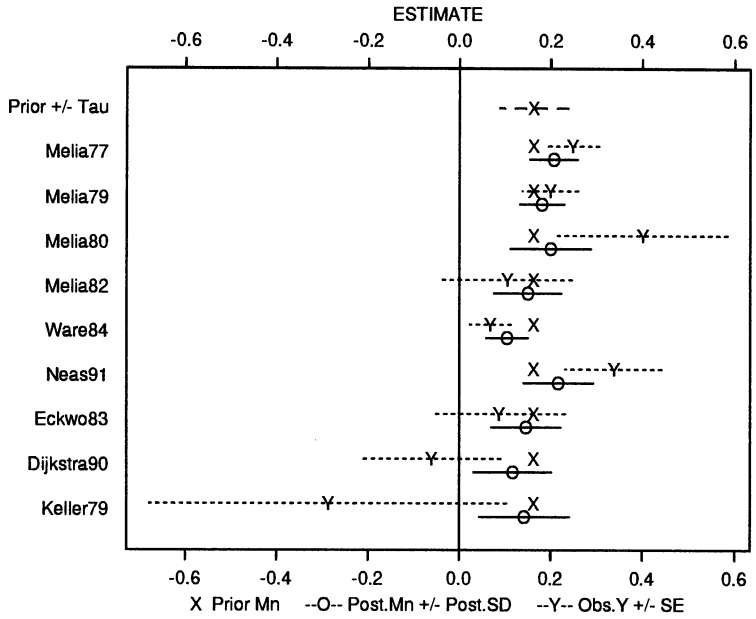

Exhibit 4. Summary graph of the NO_2 meta-analysis comparing the single-study estimates and standard errors with the prior and posterior means and standard deviations

Below this error bar are sets of error bars, one pair for each study, that compare the single-study results with the posterior distributions for each effect. Plotted with a "Y" are the values of y_i, with horizontal (dotted line) error bars extending out s_i in each direction. Plotted with an "O" and a solid error bar are the values of $\theta_i^* \pm \theta_i^{**1/2}$. The two error bars are slightly offset vertically to avoid overprinting, and the "X" denoting the location of μ^* is repeated for each study. This provides a visualization of how the Bayesian computations convert the set $\{y_i \pm s_i\}$ into the set $\{\theta_i^* \pm \theta_i^{**1/2}\}$. Each "O" is an average of an "X" and a "Y". Exhibit 4 shows the effect of "borrowing strength", in that the posterior standard deviations are less, much less for most studies, than the within-study standard errors. However, this is not always true. In some cases uncertainty in τ leads to uncertainty in how much to shrink and results in $\theta_i^{**1/2} > s_i$ for one or more studies.

2.4 The prior distribution for τ

More than any other feature, the analyses for combining information advocated here differ from the other commonly used approaches in the use of a proper prior distribution for τ, the computation of its posterior distribution, and averaging with respect to this posterior distribution to obtain inferences regarding all other parameters. The crucial formulas (8-11) require the entire *distribution* of τ, rather than a single value, known or estimated. The previous example showed how this distribution, depicted by the histogram in the trace plot of the analysis, plays a key role in the combining-information analysis. If μ has a diffuse prior distribution, the posterior distribution of τ is given in the Appendix as

$$\pi(\tau \mid y) \propto \pi(\tau) \int \Pi_i (\tau^2+s_i^2)^{-1/2}\exp\{-[y_i-\mu]^2/2(\tau^2+s_i^2)\}\, d\mu \tag{12}$$

$$\propto \pi(\tau)\, (\Sigma_j(\tau^2+s_i^2)^{-1})^{-1/2}\, \Pi_i (\tau^2+s_i^2)^{-1/2}\exp\{-[y_i-\mu^*(\tau)]^2/2(\tau^2+s_i^2)\}$$

The posterior distribution is proportional to the product of the prior distribution and the other factors on the right side of (12), the integrated likelihood (also called the restricted likelihood) of τ. If K, the number of values of i, is very large, and if the prior $\pi(\tau)$ is very highly dispersed, then this second factor will determine the shape of $\pi(\tau \mid y)$ and the exact shape of $\pi(\tau)$ will not matter much. This is called the *principle of stable estimation* (Edwards, Lindman and Savage, 1963). However, in many combining-information applications K is small, and it is important to realize then that the choice of distribution $\pi(\tau)$ may critically affect the analysis. This is not a weakness of the Bayesian approach but merely a reflection of the true uncertainties inherent in the problem of combining information from diverse sources. For example, the fixed-effect approach to combining information, which assumes that every $\theta_i = \mu$, is equivalent to assuming that $\pi(\tau)$ is concentrated at or very near $\tau = 0$, and any justification for applying the fixed-effect model in a given situation is exactly equivalent to justifying the use of such a prior distribution for τ.

The use of a prior distribution for τ is best thought of as a compromise between opposing philosophies about meta-analysis: those who believe that τ is near 0 (the philosophy of a fixed effect meta-analysis) and those who believe that τ is large and borrowing strength is hopeless in most cases (the "you can't combine apples and oranges" philosophy). An open-minded prior distribution should assign significant prior probabilities that either school could be right for any given problem.

As stated earlier, the prior density for τ used in the previous example is

$$\pi(\tau) = s_0 /(s_0 + \tau)^2 \tag{13}$$

where

$$s_0^2 = K / \Sigma s_i^{-2} \tag{14}$$

Although this particular choice has no claim to being optimal, it does have the following desirable properties

1. The density has a maximum at $\tau = 0$ and is a decreasing function of τ. This is desirable since the very concept of being able to combine the values of y_i focuses on the possibility $\tau = 0$, i.e., that every y_i has the same mean.

2. The density is positive for all $\tau \geq 0$ and integrates to 1, making it a proper density.

3. The density is very highly dispersed, since the expectation of both τ and $1/\tau$ are infinity. The density is "open-minded" toward both very small and very large values of τ. The 1st and 99th percentiles of τ are $s_0/99$ and $99s_0$, respectively.

4. The quartiles of the distribution are $s_0/3$, s_0, and $3s_0$, where s_0 is a "typical" value of s_i, based on the harmonic mean of the sample variances. If the s_i are not all equal, the definition (14) of s_0 tilts the average toward the smaller s_i, whose values of y_i are more informative about τ. It is desirable to maintain comparability between the two sources of variation in the model, represented by τ and s, and the use of the harmonic mean, rather than the arithmetic mean, of the sampling variances prevents the analysis from being unduly influenced by the addition of one or two very imprecise y_i having large s_i.

5. Let $B(\tau, s) = \tau^2/(\tau^2 + s^2)$ be the shrinkage factor [cf. eq. (7)] for a study having standard deviation s if τ is known. Then the prior quartiles of $B(\tau, s_0)$ are .1, .5 and .9, exemplifying the open-mindedness inherent in this choice of $\pi(\tau)$. There is a 25% probability of virtually complete shrinkage of the typical y (having $s = s_0$), namely $B < .1$, while there is also 25% probability of virtually no shrinkage, namely $B > .9$, so that the opinions of both ends of the spectrum, extreme believers and nonbelievers in the possibility of borrowing strength, are well represented. The prior distribution of $B(\tau, s_0)$ is symmetric about the value $B = .5$ with a U-shaped density.

No doubt many highly dispersed distributions would have properties similar to 1-5 above, and would work about as well as a default distribution for τ. The improper prior distribution $\pi(\tau) \propto 1/\tau$ is sometimes suggested in Bayesian applications involving an unknown standard deviation. That choice cannot be used here because it leads to a non integrable posterior distribution for τ. Another commonly suggested "noninformative" distribution, the uniform diffuse prior, $\pi(\tau) = 1$ for $\tau \geq 0$, does produce an integrable posterior distribution, but it lacks the desirable properties 1 and 2 above, and thus tends to skew the posterior distribution toward large values of τ.

Sensitivity Analysis for $\pi(\tau)$

The prior distribution (13) can be viewed as one choice from a family of distributions defined by

$$\pi(\tau; \tau_0) = \tau_0 /(\tau_0 + \tau)^2 \qquad (15)$$

where the choice $\tau_0 = s_0$ is viewed as being "neutral" about the question of how much to shrink the estimates toward a common value. At one extreme, choosing $\tau_0 << s_0$ weights the analysis toward the conclusion of a fixed-effect model that assumes all θ_i are equal, while at the other extreme choosing $\tau_0 >> s_0$ produces a result in which there is relatively little "borrowing strength." The latter situation also leads to a large value of μ^{**}, the posterior variance of μ, agreeing with those who believe that not much benefit is to be gained by combining the y_i in a single analysis. The question of how small or how large

τ_0 needs to be in order to produce these extreme results depends on the size of K. If few studies are being combined, the value of τ_0 is more influential than if K is large.

For example, choosing $\tau_0 = s_0/3$ is equivalent to being willing to bet even odds that the shrinkage factor $B(\tau, s_0)$ is above or below 0.1, while choosing $\tau_0 = 3s_0$ asserts that this apriori median is 0.9. Assuming that s_0 defined by (14) is a "typical" value of s_i, it is not recommended to choose τ_0 outside the range $s_0/3 < \tau_0 < 3s_0$ unless there is extensive prior knowledge about the distribution of the θ_i across studies. These represent a range of extremely different prior beliefs, but the differences in the corresponding posterior distributions are much less. Exhibit 5 shows what happens when these different priors are applied to the NO_2 example data. Because the prior density has such heavy tails, scaling the distribution by a factor of 3 in either direction did not make much inferential difference. As Exhibit 5 shows, the posterior moments of τ are affected the most, while those of μ are essentially unchanged. The posterior distributions of the θ_i are more sensitive to the prior as s_i increases. In Exhibit 5 only results pertaining to θ_{KI79}, which has the largest s_i, are shown, and the effects of changing τ_0 by a factor of 9 are not dramatic -- e.g. $P(\theta_{KI79} > 0)$ goes from 0.96 to 0.90.

τ_0	τ	μ	θ_{KI79}	$P(\theta_{KI79} > 0)$
0.027	0.056 (0.046)	0.162 (0.041)	0.148 (0.081)	0.957
0.082	0.075 (0.051)	0.163 (0.046)	0.142 (0.098)	0.929
0.246	0.091 (0.057)	0.163 (0.051)	0.134 (0.113)	0.899

Exhibit 5. Posterior means (posterior standard deviations) of parameters in the NO_2 meta-analysis as a function of τ_0, the prior median of τ, assuming $\pi(\tau) = \pi(\tau; \tau_0)$ of eq. (15).

3. Making Use of Study-Level Covariates

3.1 The Bayesian Regression Model

We next consider more sophisticated meta-analyses, in which characteristics of the studies are used to model differences between studies. In the analyses combining the y_i, three sources of variation are considered:

a) between-study differences explained by fixed characteristics of the studies,

b) unexplained random variation from study to study, and

c) random sampling error within each study.

The central task of a meta-analysis is to allocate variation in y_i among these three sources, and a combining-information analysis must estimate and report on all three, including appropriate uncertainty assessments.

The three sources of variation correspond to the three terms in the following model for the y_i,

$$y_i = \boldsymbol{x}_i\beta + \delta_i + \varepsilon_i \tag{16a}$$

$$\theta_i = \boldsymbol{x}_i\beta + \delta_i \tag{16b}$$

$$\delta_i \sim N(0, \tau^2) \tag{16c}$$

$$\varepsilon_i \sim N(0, s_i^2) \tag{16d}$$

The term $\boldsymbol{x}_i\beta$ in (16a) represents source a) above, the predictable part of the variation in y. Here $\boldsymbol{x}_i$ denotes the row vector $(1, x_{i1}, ..., x_{iJ})$, where the J study characteristics x_{ij}, $(j = 1, ..., J)$ are used to explain study differences and the parameter column vector $\beta = (\beta_0, \beta_1, ..., \beta_J)^t$ are coefficients that need to be estimated. The matrix multiplication representation

$$\boldsymbol{x}_i\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_J x_{iJ}$$

shows that this term has the form of a typical regression equation. Model (16) is a straightforward generalization of (1), where μ is replaced by $\boldsymbol{x}_i\beta$. The discussion paper by DuMouchel and Harris (1983) was one of the first to apply this model to meta-analysis. They used this model to combine the results of 37 dose-response studies in humans and other species, as a way of attacking the species extrapolation problem in risk assessment.

Since the values of s_i are assumed known, the allocation of variation in the y_i is achieved by estimating the unknown parameters in the other two terms, namely, by estimating β and τ.

The size of τ measures how much artifactual study-specific factors like population variations, investigator differences, uncontrolled conditions of each study, and so forth, contribute to unpredictable deviations from the regression prediction $\boldsymbol{x}_i\beta$. These deviations are unpredictable because no x-variables have been identified that isolate them, and they are not reduced in size or importance as the size of the study increases. If and when such an x-variable is identified, putting it into the $\boldsymbol{x}_i\beta$ term would tend to reduce the value of τ in the revised model and correspondingly reduce the role of the δ_i.

Prior and Posterior Distributions

The Bayesian model (16) is completed by assigning prior distributions to β and τ. Analogous to (2),

$$\beta_j \sim N(m_j, d_j^2) \qquad j = 0, 1, ..., J \quad (17a)$$

$$\tau \sim \pi(\tau) \quad (17b)$$

where, usually, each $d_j \to \infty$ to represent the situation of no prior knowledge about β. (However, DuMouchel and Harris (1983) and DuMouchel and Groër (1989) showed how informative prior distributions for some elements of β can be used to insert scientific knowledge into the analysis. These papers use a slightly more general formulation in which (17a) is replaced by the assumption that the vector β has a multivariate normal prior distribution with known mean vector and covariance matrix.)

The Bayesian calculations when covariates are present are straightforward generalizations of the case without covariates. As detailed in DuMouchel and Harris (1983) and DuMouchel (1990), for each fixed τ standard weighted regression computations lead to multivariate normal posterior distributions for β and θ. The REML likelihood function $L(\tau)$ is multiplied by $\pi(\tau)$ to produce $\pi(\tau \mid y)$. Finally, the marginal moments and distributions of β and θ are computed by averaging the distributions conditional on τ with respect to $\pi(\tau \mid y)$.

3.2 Example: Use of Covariates Describing the NO_2 Studies

Returning to the nine studies of the effects of NO_2 on childhood respiratory illness, we consider how the design of these studies differed. Hasselblad et al (1992) discuss these differences in detail. For the purpose of this example, we consider just whether the reported odds ratio was computed from an analysis that:

a) adjusts for the smoking status of the parents

b) uses a measured NO_2 dose instead of a surrogate (presence or absence of a gas stove)

c) adjusts for the gender of the child

In principle, we would prefer that every study did all three of the above. In fact, only one did. Exhibit 6 lists the data from the nine studies, including whether (Y) or not (N) each study used the preferred methodology on each of the three issues labeled "Smoke", "NO2" and "Gender" in Exhibit 6.

The primary outcome for each study is the odds ratio (OR in Exhibit 6) for childhood respiratory illness at a specific level of NO_2 concentration in the home. The odds ratio for each study was computed from a logistic regression of children's illness on NO_2 exposure, with additional terms adjusting for various possible confounders, different from study to study. For example, Exhibit 6 shows that only two of the nine studies measured and adjusted for whether parents of the children smoked (Smoke = "Y" in Table 3), while five of the studies adjusted for the gender of the child whose respiratory illness was being evaluated (Gender = "Y"), and only 4 of the studies (NO2 = "Y") actually measured NO_2 concentrations in homes; the others merely used a dummy variable for the presence of a surrogate for high NO_2 (a gas stove) and imputed the NO_2 concentrations for the analyses. Presumably, these quite different study designs could easily affect their estimated odds ratios, since the meaning of a logistic regression coefficient varies, depending on what other variables are in the regression. A hierarchical model can attempt to estimate and adjust for such differences.

In order to define a term $\boldsymbol{x}\beta$ to describe each study's design, let a numerical version of the variables *Smoke*, *NO2* and *Gender* take the value 1 if "N" and 0 if "Y". That is, study i has the value $Smoke_i = 1$ if its log odds ratio is potentially biased because it did not adjust for household smoking (otherwise $Smoke_i = 0$), with analogous definitions for the variables *NO2* and *Gender*. Now define

$$\boldsymbol{x}_i\beta = \beta_0 + \beta_1 Smoke_i + \beta_2 NO2_i + \beta_3 Gender_i \tag{18}$$

The parameter β_0 is interpreted as the average log odds ratio of all studies designed to handle all three of the above problems and is the primary parameter of interest in the meta-analysis. The parameter β_1 is defined as the average bias in the log odds ratio (NO_2 vs. childhood respiratory illness) among studies that do not adjust for household smoking, the parameter β_2 is defined as the average bias in estimation of that log odds ratio among studies that use a surrogate measure of NO_2 exposure, and β_3 is defined similarly with respect to the Gender adjustment characteristic. Furthermore, we assume that among studies that fail to adjust for more that one of these issues, these biases add.

That is, the response y_i is an estimated logistic regression coefficient and we model its variation using the specification of terms in its corresponding model, adding also a study-specific random effect and sampling error. Thus, letting y_i denote the values of *ln.or* in Exhibit 6, we assume that θ_i is the mean of y_i in the context of study i, but that θ_i has the distribution $N(\boldsymbol{x}_i\beta, \tau^2)$, where $\boldsymbol{x}_i\beta$ is given by (18) and τ is the standard deviation of any extra variation in θ not explained by (18).

```
> NO2.frame
           Smoke NO2 Gender   OR Lower Upper  ln.OR  se.lnOR
   Melia77     N   N      Y 1.28  1.14  1.43  0.247   0.058
   Melia79     N   N      Y 1.22  1.08  1.37  0.199   0.061
   Melia80     N   Y      Y 1.49  1.04  2.14  0.399   0.184
   Melia82     N   Y      Y 1.11  0.84  1.46  0.104   0.141
    Ware84     N   N      N 1.07  0.98  1.17  0.068   0.045
    Neas91     Y   Y      Y 1.40  1.14  1.72  0.336   0.105
   Eckwo83     Y   N      N 1.09  0.82  1.45  0.086   0.145
Dijkstra90     N   Y      N 0.94  0.70  1.27 -0.062   0.152
  Keller79     N   N      N 0.75  0.35  1.62 -0.288   0.391
```

Exhibit 6. Data describing the NO_2 studies (from Hasselblad et al 1992 and Kotchmar 1993)

Exhibit 7 shows a portion of an interactive computer session performing the Bayesian regression analysis of these data. The program "hblm" (hierarchical Bayesian linear model) is written in the language S-PLUS® (Statistical Science, 1993).

```
> NO2.hb <- hblm(ln.OR~Smoke+NO2+Gender, s=se.lnOR)
> NO2.hb                                   {All analysis results bundled in the variable NO2.hb}

Coefficients:
             Mean    S.D.  Prob > 0
(Intercept)  0.173  0.054  0.997        {Estimate of the effect if a study}
      Smoke -0.050  0.057  0.180        {adjusts for all confounders}
        NO2  0.012  0.049  0.606
     Gender -0.097  0.040  0.007        {Estimate of the bias if Gender is ignored}

      RSS Estimate of Tau =  0          {Note that τ is estimated to be small}
      Prior Median of Tau =  0.082
    Posterior Mean of Tau =  0.043    (s.d. =  0.046 )

> plot(NO2.hb$trace)                   {produces Exhibit 8}
> plot(NO2.hb)                         {produces Exhibit 9}
```

Exhibit 7. HBLM analysis input (bold) and partial output (unbold) for the NO_2 meta-analysis with covariates (with annotations at right in { })

The line labeled (Intercept) in Exhibit 7 shows that β_0 is estimated to be 0.173 ± 0.054, $P(\beta_0 > 0 \mid y) = 0.997$, which is in close agreement with the values $\mu^* = 0.163 \pm 0.046$, $P(\mu > 0 \mid y) = 0.998$ from the previous meta-analysis without covariates. The estimated biases in the log odds ratio due to neglecting the issues of smoking, NO_2 dose measurement, and gender of child are -0.050, 0.012 and -0.097, respectively, although only the bias due to neglecting gender is significantly different from zero.

The last three output lines in Exhibit 7 show information about the estimation of τ. With this model, a classical point estimate [based on a residual sum of squares, see DuMouchel and Harris (1983)] of τ is 0, but the posterior distribution $\pi(\tau \mid y)$ has mean 0.043 and standard deviation 0.046, compared to $s_0 = 0.082$. The use of covariates in the meta-analysis has reduced the posterior mean of τ from 0.075, shown in Exhibit 5, to 0.043. Because the REML point estimate of τ is 0, both the fixed and non Bayesian mixed effects methods would compute the estimates and standard errors of β as if τ were known to be 0. Separate calculations, not shown in Exhibit 7, show that the resulting one-sided P-values for the 4

coefficients would be 0.9999, 0.16, 0.62 and 0.0007, respectively — such an analysis would exaggerate the significances of β_0 and β_3 by factors of 30 and 10, respectively.

Exhibit 8 shows the trace plot for the meta-analysis with covariates. Compared to Exhibit 3, the histogram of $\pi(\tau \mid y)$ is shifted to the left. When covariates are present, the estimates $\theta_i^*(\tau)$ do not shrink to a common value as τ approaches 0 but shrink toward the estimates of $x_i\beta$. Studies having the same values of *Smoke*, *NO2* and *Gender* shrink towards each other, but studies whose regression fitted values are far apart remain far apart after shrinking. Thus, for example, the outlying study Kl79, labeled "J" in Exhibits 3 and 8, shrinks less in the latter analysis.

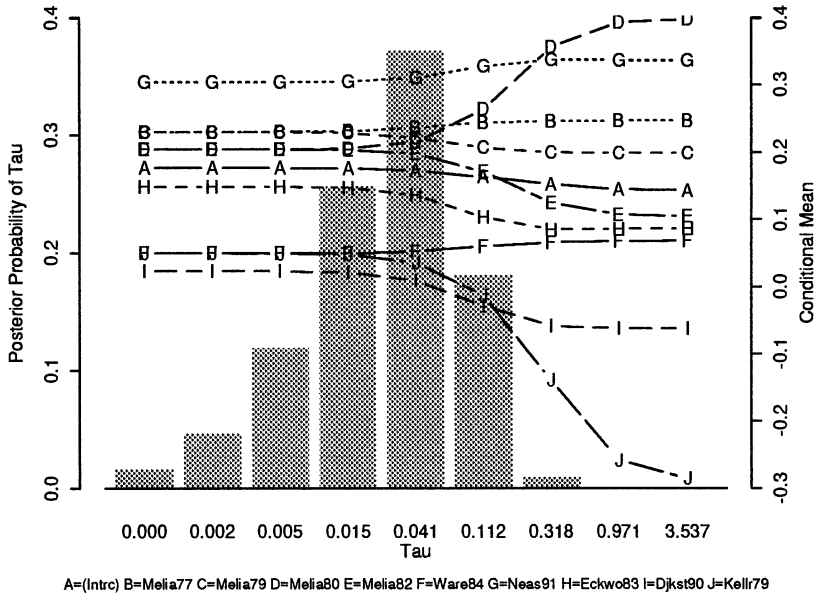

Exhibit 8. Trace plot for the NO_2 meta-analysis with covariates

This is also shown in Exhibit 9, the summary plot of the meta-analysis with covariates, and may be contrasted with Exhibit 4. In Exhibit 9, the values of $x_i\beta$ are plotted as "X" on the horizontal scale, and it may be seen that the three studies that did not adjust for either smoking or gender (Ware84, Dijkstra90 and Keller79) have much smaller posterior means than in the previous analysis. There is now no certainty at all that a repeat of the last two studies with a very large sample size would lead to a positive log odds ratio, since the posterior error bars for these studies cross 0 on the graph.

This example has shown how the Bayesian approach can make good use of measures of quality of studies. The three variables *Smoke*, *NO2*, and *Gender* are quality measures ("Y" is good, "N" is poor) denoted generically by x. The analysis estimates the relationship between study outcome y and x; if there is no relationship then there is no advantage to downweighting the poorer quality studies. If, as in this example, x is coded as 0 for the highest quality study, then the intercept β_0 will be the parameter of

interest, and studies having x near 0 will automatically carry greater weight for estimation of the intercept than will studies having larger values of x.

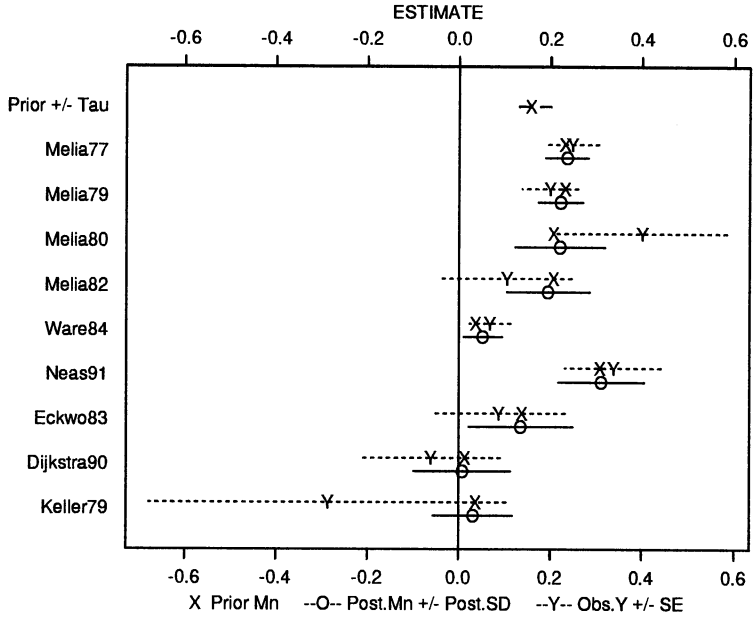

Exhibit 9. Summary plot for the NO_2 meta-analysis with covariates

3.3 Summary and Discussion

Compared to simpler methods of combining information across studies, the hierarchical Bayesian linear model has the advantage of allowing for more sources of variation in the data. It draws on the familiar statistical methods of regression and mixed-model analysis. This allows the use of graphical and model-checking techniques like cross-validation and the analysis of residuals in a meta-analytic context, as in DuMouchel (1994). A judicious choice of study-level covariates, when available, helps reduce the unexplainable variation between studies and may lead to new substantive insights. The presence of the between-studies variance component is a complication compared to a standard regression analysis, and the trace plot (Exhibits 2, 3 and 8) is a useful graph for explaining the role of τ in the analysis. A summary plot (Exhibits 4 and 9) is useful for comparing the individual study results to the hierarchical modeling results. The Bayesian method allows the input of a prior for τ if it is desired to "tilt" the analysis toward or away from a fixed-effect model, and of a prior for β if it is desired to input scientific knowledge regarding the expected effect of some of the covariates, as in DuMouchel and Harris (1983).

The fact that the Bayesian calculations average over a full range of values of τ, even if its point estimate is 0, make this method more stable when only a few studies are being combined. Classical confidence

intervals tend to be short if study heterogeneity is not significant, and much longer if it is, while the corresponding Bayesian intervals will be of an intermediate length most of the time.

Another advantage of the hierarchical approach is that it encourages thinking about ensembles of studies rather than single studies. When researchers design new studies of a scientific problem, it is advantageous to plan how the new study could best complement existing studies. If a hierarchical model relating different studies has been formed, the theory of design of experiments can be used to choose a next study that leads to a more powerful future hierarchical meta-analysis.

There are many avenues open for future research in this area. One important extension of the model is to relax the assumption of normality of the random effects δ_i. For example, assumption of a higher-tailed distribution, such as a t-distribution with very few degrees of freedom, would allow outlier studies to have less influence on the estimate of the superpopulation mean. Other more complex specifications of the prior distributions can allow the fitting of clusters or groupings of the θ_i, even when no corresponding x-variables have been identified. See Malec and Sedransk (1992) for an example.

Another desireable extension of the methodology would allow each individual summary y_i to be a vector. For example, the relative risk of an exposed population with respcct to several diseases or types of cancer. To take another example, specifying nonlinear dose-response curves usually requires more than one parameter, so each of several dose-response studies might report a vector of parameter estimates, with an estimated covariance matrix. The random effects would also be vectors, and the parameter τ would now be a matrix, for which the Bayesian methodology would require a prior distribution and an integration scheme. Such a statistical model would be formally similar to the longitudinal random effects models of Laird and Ware (1982).

The model fitting and graphical techniques discussed in this paper have been implemented in the statistical computing environment S-PLUS® (Statistical Science 1993). For information on obtaining the programs, please email the author at `dumouch@bayes.cpmc.columbia.edu`. Other Bayesian software packages, like the BUGS package (Spiegelhalter et al, 1994) can also fit hblm models. As discussed in DuMouchel and Waternaux (1994) or Bryk and Raudenbush (1992) general-purpose mixed model software, such as PROC MIXED (SAS Institute 1992) can fit an empirical Bayes version of this meta-analytic model.

References

Bryk A and Raudenbush S, 1992, *Hierarchical Linear Models.* Sage Publications. Newbury Park, CA.

Chalmers TC, et al (1987a) Meta-analysis of clinical trials as a scientific discipline, I: Control of bias and comparison with large cooperative trials. *Statistics in Medicine* 6: 315-325.

Chalmers TC, et al (1987b) Meta-analysis of clinical trials as a scientific discipline, II: Replicate variability and comparison of studies that agree and disagree. *Statistics in Medicine* 6: 733-744.

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

DuMouchel W (1990) Bayesian Meta-analysis. In *Statistical Methodology in Pharmaceutical Sciences,* D. Berry, Ed., Marcel Dekker, 1990

DuMouchel W (1994) Predictive Crossvalidation of Bayesian Meta-analyses, in *Proceedings of Fifth Valencia International Meeting on Bayesian Statistics,* J. Bernardo et al, eds., Valencia, Spain.

DuMouchel W, Groër PG (1989) A Bayesian Methodology for Scaling Radiation Studies from Animals to Man. *Health Physics* 57: 411-418.

DuMouchel W, Harris J (1983) Bayes methods for combining the results of cancer studies in humans and other species. *J. American Statistical Assoc.*, 78:293.

DuMouchel W, Waternaux C (1994) Bayesian meta-analyses with general-purpose software, in *Proceedings of Interface '94*, John Sall, ed., Research Triangle Park, NC: Interface Foundation.

Edwards WL, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research, *Psychological Review* 70:193-242.

Gaver D, Draper D, Goel P, Greenhouse J, Hedges L, Morris M, Waternaux C (1992) *Combining Information: Statistical Issues and Opportunities for Research,* National Academy Press, Washington, DC.

Hasselblad V, Eddy DM, Kotchmar DJ (1992) Synthesis of environmental evidence: nitrogen dioxide epidemiology studies, *J. Air & Waste Management Assoc.* 42:662-71.

Hedges L, Olkin I (1985) *Statistical Methods for Meta Analysis,* Academic Press, Orlando, FL.

Kotchmar DJ (1993) Synthesis of environmental evidence: nitrogen dioxide epidemiology studies, Table 1 from handout at oral presentation, US EPA/NISS Workshop on Statistical Methods for Combining Environmental Information, September 27-28, 1993, Chapel Hill, NC.

Laird NM, Ware JH (1982) Random-effects models for longitudinal data, *Biometrics*, 38:963-974.

Light RL, Pillemer DB (1984) *Summing Up: the Science of Reviewing Research*, Cambridge, MA: Harvard University Press.

Louis TA, Fineberg HV, Mosteller F (1985) Findings for public health from meta-analysis, *Annual Review of Public Health* 6:1-20.

Malec D, Sedransk J (1992) Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain, *Biometrika* 79:593-601.

Morris CN, Normand SL (1992) Hierarchical Models for Combining Information and for Meta-analysis, in *Bayesian Statistics 4,* pp.321-344. Oxford University Press.

Peto R (1987) Discussion during the workshop on methodological issues in overviews of randomized clinical trials, *Statistics in Medicine*, 6:229. [The entire May 1987 issue of *Statistics in Medicine* is devoted to meta-analysis.]

Rosenthal R (1979) The "file drawer problem" and the tolerance for null results, *Psychological Bulletin* 66:638-641.

Sacks HS, Berrier J, Reitman D, Ancona Berk, VA, Chalmers TC (1987) Meta-analyses of randomized clinical trials. *New England J. Medicine* 316:450-455.

SAS Institute Inc. (1992) SAS Technical Report P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute, Inc., 620 pp.

Spiegelhalter DJ, Thomas A, Best NG (1994) Computation on Bayesian graphical models, in *Proceedings of Fifth Valencia International Meeting on Bayesian Statistics,* J. Bernardo et al, eds., Valencia, Spain.

Statistical Sciences, Inc. (1993) *S-PLUS Users Manual*, Seattle: Statistical Sciences, Inc.

Appendix

This Appendix collects together the formulas for a Bayesian hierarchical meta-analysis without covariates and with a diffuse prior for μ. For the model with covariates, see DuMouchel and Harris (1983) or DuMouchel (1990).

$$y_i = \mu + \delta_i + \varepsilon_i \tag{1a}$$

$$\theta_i = \mu + \delta_i \tag{1b}$$

$$\delta_i \sim N(0, \tau^2) \tag{1c}$$

$$\varepsilon_i \sim N(0, s_i^2) \tag{1d}$$

$$\mu \sim N(m, d^2 \to \infty) \tag{2a}$$

$$\tau \sim \pi(\tau) \tag{2b}$$

$$\pi(\tau) = s_0 /(s_0 + \tau)^2 \tag{3}$$

$$s_0^2 = K / \Sigma s_i^{-2} \tag{4}$$

$$\mu^*(\tau) \equiv E[\mu \mid \boldsymbol{y}, \tau] = \Sigma_i \, w_i(\tau) \, y_i \tag{5}$$

$$w_i(\tau) = (\tau^2 + s_i^2)^{-1} / \Sigma_j \, (\tau^2 + s_j^2)^{-1} \tag{6}$$

$$V[\mu \mid \boldsymbol{y}, \tau] = 1/\Sigma_j(\tau^2 + s_j^2)^{-1}$$

$$\theta_i^*(\tau) \equiv E[\theta_i \mid \boldsymbol{y}, \tau] = \mu^*(\tau) + [y_i - \mu^*(\tau)] \, \tau^2/(\tau^2 + s_i^2) \tag{7}$$

$$V[\theta_i \mid \boldsymbol{y}, \tau] = [s_i^2/(\tau^2 + s_i^2)]^2/\Sigma_j(\tau^2 + s_j^2)^{-1} \; + \tau^2 s_i^2/(\tau^2 + s_i^2)$$

$$\pi(\tau \mid \boldsymbol{y}) \propto \pi(\tau) \int \Pi_i \, (\tau^2+s_i^2)^{-1/2} \exp\{-[y_i-\mu]^2/2(\tau^2+s_i^2)\} \, d\mu \tag{12}$$

$$\propto \pi(\tau) \, (\Sigma_j(\tau^2+s_i^2)^{-1})^{-1/2} \, \Pi_i \, (\tau^2+s_i^2)^{-1/2} \exp\{-[y_i-\mu^*(\tau)]^2/2(\tau^2+s_i^2)\}$$

$$\mu^* \equiv E[\mu \mid \boldsymbol{y}] = \int \mu^*(\tau) \, \pi(\tau \mid \boldsymbol{y}) \, d\tau \tag{8}$$

$$\mu^{**} \equiv V[\mu \mid \boldsymbol{y}] = \int \{V[\mu \mid \boldsymbol{y}, \tau] + [\mu^*(\tau) - \mu^*]^2\} \, \pi(\tau \mid \boldsymbol{y}) \, d\tau \tag{9}$$

$$P(\mu > 0 \mid \boldsymbol{y}) = \int \Phi(\mu^*(\tau)/V[\mu \mid \boldsymbol{y}, \tau]^{1/2}) \, \pi(\tau \mid \boldsymbol{y}) \, d\tau \tag{10}$$

$$\theta_i^* \equiv E[\theta_i \mid \boldsymbol{y}] = \int \{\mu^*(\tau) + [y_i - \mu^*(\tau)] \, \tau^2/(\tau^2 + s_i^2)\} \pi(\tau \mid \boldsymbol{y}) \, d\tau \tag{11}$$

$$\theta_i^{**} \equiv V[\theta_i \mid \boldsymbol{y}] = \int \{V[\theta_i \mid \boldsymbol{y}, \tau] + [\theta_i^*(\tau) - \theta_i^*]^2\} \, \pi(\tau \mid \boldsymbol{y}) \, d\tau$$