A Bayesian Semi-Parametric Model for Random Effects Meta-Analysis

Deborah Burr
School of Public Health
Ohio State University
Columbus, OH 43210

Hani Doss
Department of Statistics
Ohio State University
Columbus, OH 43210

Revised, June 2004

Abstract

In meta-analysis there is an increasing trend to explicitly acknowledge the presence of study variability through random effects models. That is, one assumes that for each study, there is a study-specific effect and one is observing an estimate of this latent variable. In a random effects model, one assumes that these study-specific effects come from some distribution, and one can estimate the parameters of this distribution, as well as the study-specific effects themselves. This distribution is most often modelled through a parametric family, usually a family of normal distributions. The advantage of using a normal distribution is that the mean parameter plays an important role, and much of the focus is on determining whether or not this mean is 0. For example, it may be easier to justify funding further studies if it is determined that this mean is not 0. Typically, this normality assumption is made for the sake of convenience, rather than from some theoretical justification, and may not actually hold. We present a Bayesian model in which the distribution of the study-specific effects is modelled through a certain class of nonparametric priors. These priors can be designed to concentrate most of their mass around the family of normal distributions, but still allow for any other distribution. The priors involve a univariate parameter that plays the role of the mean parameter in the normal model, and they give rise to robust inference about this parameter. We present a Markov chain algorithm for estimating the posterior distributions under the model. Finally, we give two illustrations of the use of the model.

1 Introduction

The following situation arises frequently in medical studies. Each of m centers reports the outcome of a study that investigates the same medical issue, which for the sake of concreteness we will think of as being a comparison between a new and an old treatment. The results are inconsistent, with some studies being favorable to the new treatment, while others indicate less promise, and one would like to arrive at an overall conclusion regarding the benefits of the new treatment.

Early work in meta-analysis involved pooling of effect-size estimates or combining of p-values. However, because the centers may differ in their patient pool (e.g. overall health level, age, genetic makeup) or the quality of the health care they provide, it is now widely recognized that it is important to explicitly deal with the heterogeneity of the studies through random effects models, in which for each center i there is a center-specific "true effect," represented by a parameter ψ_i.

Suppose that for each i, center i gathers data D_i from a distribution $P_i(\psi_i)$. This distribution depends on ψ_i and also on other quantities, for example the sample size as well as nuisance parameters specific to the i^{th} center. For instance, ψ_i might be the regression coefficient for the indicator of treatment in a Cox model, and D_i is the estimate of this parameter. As another example, ψ_i might be the ratio of the survival probabilities at a fixed time for the new and old treatments, and D_i is a ratio of estimates of survival probabilities based on censored data. A third example, which is very common in epidemiological studies, is one in which ψ_i is the odds ratio arising in case-control studies, and D_i is either an adjusted odds ratio based on a logistic regression model that involves relevant covariates, or simply the usual odds ratio based on a 2×2 table.

A very commonly used random effects model for dealing with this kind of situation is the following:

$$\text{Conditional on } \psi_i, \quad D_i \overset{\text{ind}}{\sim} \mathcal{N}(\psi_i, \sigma_i^2), \quad i = 1, \ldots, m \tag{1.1a}$$

$$\psi_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \quad i = 1, \ldots, m \tag{1.1b}$$

In (1.1b), μ and τ are unknown parameters. (The σ_i's are also unknown, but we will usually have estimates $\hat{\sigma}_i$ along with the D_i's, and estimation of the σ_i's is secondary. This is discussed further in Section 2).

Model (1.1) has been considered extensively in the meta-analysis literature, with much of the work focused on the case where ψ_i is the difference between two binomial probabilities or an odds ratio based on two binomial probabilities. In the frequentist setting, the classical paper by DerSimonian and Laird (1986) gives formulas for the maximum likelihood estimates of μ and τ, and a test for the null hypothesis that $\mu = 0$. In a Bayesian analysis, a joint prior is put on the pair (μ, τ). Bayesian approaches are developed in a number of papers, including Skene and Wakefield (1990), DuMouchel (1990), Morris and Normand (1992), Carlin (1992), and Smith et al. (1995). As is discussed in these papers, a key advantage of the Bayesian approach is that inference concerning the center-specific effects ψ_i is carried out in a natural manner through consideration of the posterior distributions of these parameters. (When Markov chain Monte Carlo is used, estimates of these posterior distributions typically arise as part of the output. From the frequentist perspective, estimation of the study-specific effects ψ_i is much

more difficult. An important application arises in small area estimation; see Ghosh and Rao (1994) for a review.)

The approximation of $P_i(\psi_i)$ by a normal distribution in (1.1a) is typically supported by some theoretical result, for example the asymptotic normality of maximum likelihood estimates. By contrast, the normality statement in (1.1b) is a modelling assumption, which generally is made for the sake of convenience and does not have any theoretical justification. One would like to replace (1.1) with a model of the following sort:

$$\text{Conditional on } \psi_i, \quad D_i \overset{\text{ind}}{\sim} \mathcal{N}(\psi_i, \sigma_i^2), \quad i = 1, \ldots, m \tag{1.2a}$$

$$\psi_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, m \tag{1.2b}$$

$$F \sim \pi, \tag{1.2c}$$

where π is a "nonparametric prior."

This paper is motivated by a situation we recently encountered (Burr et al. 2003), in which we considered 12 papers appearing in medical journals, each of which reported on a case-control study that aimed to determine whether or not the presence of a certain genetic trait was associated with an increased risk of coronary heart disease. Each study considered a group of individuals with coronary heart disease and another group with no history of heart disease. The proportion having the genetic trait in each group was noted and an odds ratio calculated. The studies gave rather inconsistent results (p-values for the two-sided test that the log odds ratio is 0 ranged from $.005$ to $.999$, and the reported log odds ratios themselves ranged from $+1.06$ to $-.38$), giving rise to sharply conflicting opinions on whether or not there could exist an association between the genetic trait and susceptibility to heart disease. It was clear that the studies had different study-specific effects, and it appeared that these did not follow a normal distribution, so that it was more appropriate to use a model of the sort (1.2) than a model based on (1.1). In our analysis, in the terminology of Model (1.1), the issue of main interest was not estimation of the center-specific ψ_i's, but rather resolving the basic question of whether the overall mean μ is different from 0, since this would determine whether or not it is justified to carry out further studies.

To deal with this issue when considering a model of the form (1.2), it is necessary that the prior π in (1.2c) involve a univariate parameter that can play a role analogous to that of μ in (1.1b). The purpose of this paper is to present a simple such model based on mixtures of "conditional Dirichlet processes." The paper is organized as follows. In Section 2, we review a standard model based on mixtures of Dirichlet processes and explain its limitations in the meta-analysis setting we have in mind. In Section 3 we present the model based on mixtures of conditional Dirichlet processes, explain its rationale, and describe a Markov chain Monte Carlo algorithm for estimating the posterior distribution. We also discuss the connection between the posterior distributions under this model and the standard model. Section 4 gives two examples that illustrate various issues. The Appendix gives a proof of a likelihood ratio formula stated in Section 3.

2 A Nonparametric Bayesian Model for Random Effects

In a Bayesian version of Model (1.1) where a prior is put on the pair (μ, τ), the most common choice is the "normal / inverse gamma" prior (see e.g. Berger (1985, p. 288), and also the

description in (2.1) below), which is conjugate to the family $\mathcal{N}(\mu, \tau^2)$. For the problem where the ψ_i's are actually observed, the posterior distribution of (μ, τ) is available in closed form. In the present situation in which the ψ_i's are latent variables on which we have only partial information, there is no closed form expression for the posterior, although it is very easy to write a MCMC algorithm to estimate this posterior, for example in BUGS (Spiegelhalter et al. 1996).

A convenient choice for a nonparametric Bayesian version of this is a model based on mixtures of Dirichlet processes (Antoniak 1974), and before proceeding, we give a brief review of this class of priors. Let $H_\theta;\ \theta \in \Theta \subset \mathbb{R}^k$ be a parametric family of distributions on the real line, and let λ be a distribution on Θ. Suppose $M_\theta > 0$ for each θ, and define $\alpha_\theta = M_\theta H_\theta$. If θ is chosen from λ, and then F is chosen from $\mathcal{D}_{\alpha_\theta}$, the Dirichlet process with parameter measure α_θ (Ferguson 1973, 1974), we say that the prior on F is a mixture of Dirichlet processes (with parameter $(\{\alpha_\theta\}_{\theta\in\Theta}, \lambda)$). Although it is sometimes useful to allow M_θ to depend on θ, for the sake of clarity of exposition we will assume that M_θ does not vary with θ, and we will denote the common value by M. In this case, M can be interpreted as a precision parameter that indicates the degree of concentration of the prior on F around the parametric family $\{H_\theta;\ \theta \in \Theta\}$. In a somewhat oversimplified but nevertheless useful view of this class of priors, we think of the family $\{H_\theta;\ \theta \in \Theta\}$ as a "line" (of dimension k) in the infinite-dimensional space of cdf's, and we imagine "tubes" around this line. For large values of M, the mixture of Dirichlet processes puts most of its mass in narrow tubes, while for small values of M the prior is more diffuse.

If we take the parametric family $\{H_\theta\}$ to be the $\mathcal{N}(\mu, \tau^2)$ family and λ to be the normal / inverse gamma conjugate prior, the model is expressed hierarchically as follows:

$$\text{Conditional on } \psi_i, \qquad D_i \overset{\text{ind}}{\sim} \mathcal{N}(\psi_i, \sigma_i^2), \quad i = 1, \ldots, m \tag{2.1a}$$

$$\text{Conditional on } F, \qquad \psi_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, m \tag{2.1b}$$

$$\text{Conditional on } \mu, \tau, \qquad F \sim \mathcal{D}_{M\mathcal{N}(\mu,\tau^2)} \tag{2.1c}$$

$$\text{Conditional on } \tau, \qquad \mu \sim \mathcal{N}(c, d\tau^2) \tag{2.1d}$$

$$\gamma = 1/\tau^2 \sim \text{Gamma}(a, b) \tag{2.1e}$$

In (2.1d) and (2.1e), $a, b, d > 0$, and $-\infty < c < \infty$ are arbitrary but fixed. The σ_i's are unknown and it is common to use estimates $\hat{\sigma}_i$ instead. If the studies do not involve small samples, this substitution has little effect (DerSimonian and Laird 1986); otherwise one will want to also put priors on the σ_i's. This kind of model has been used successfully to model random effects in many situations. An early version of the model and a Gibbs sampling algorithm for estimating posterior distributions were developed in Escobar (1988, 1994). Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)$ and $\boldsymbol{D} = (D_1, \ldots, D_m)$. In essence, the Gibbs sampler runs over the vector of latent variables $\boldsymbol{\psi}$ and the pair (μ, τ), and the result is a sample $(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)});\ g = 1, \ldots, G$, which are approximately distributed according to the conditional distribution of $(\boldsymbol{\psi}, \mu, \tau)$ given the data. See the papers in Dey et al. (1998) and Neal (2000) for recent developments concerning models of this kind.

In this paper, we will use $\mathcal{L}$ generically to denote distribution or law. We will adopt the convention that subscripting a distribution indicates conditioning. Thus, if U and V are two random variables, both $\mathcal{L}(U \mid V, \boldsymbol{D})$ and $\mathcal{L}_{\boldsymbol{D}}(U \mid V)$ mean the same thing. However, we will

use $\mathcal{L}_{\boldsymbol{D}}(U \mid V)$ when we want to focus attention on the conditioning on V. This is useful in describing steps in a Gibbs sampler, for example, when $\boldsymbol{D}$ is fixed throughout.

As mentioned earlier, in the kind of meta-analysis we have in mind, the question of principal interest is whether or not the mean of F, the distribution of the study-specific effect, is different from 0. Note that in Model (2.1), μ is not equal to $\eta = \eta(F) = \int x\, dF(x)$, the mean of F, and so inference on η is not given automatically as part of the Gibbs sampler output.

The method of Gelfand and Kottas (2002) (see also Muliere and Tardella 1998) can be used to estimate the posterior distribution of η. This method is based on Sethuraman's (1994) construction, which represents the Dirichlet process as an infinite sum of atoms, with an explicit description of their locations and sizes. In brief, the method of Gelfand and Kottas (2002) involves working with a truncated version of this infinite sum. How far out into the sum one needs to go in order to get accurate results depends on $M + m$, the parameters of the model, and the data. In particular, the truncation point needs to grow with the quantity $M + m$, and the algorithm gives good results when this quantity is small or moderate but can be quite slow when it is large. In principle, the truncation point can be chosen through informal calculations, but it is difficult to do so when implementing a MCMC algorithm because then the parameters of the Dirichlet process are perpetually changing. [We mention that we have used Sethuraman's construction to generate a Gibbs sampler in previous work (Doss 1994). The situation encountered in that paper was quite different however, in that there, we needed to generate *random variables* from an F with a Dirichlet distribution. The algorithm we used required us to generate only a segment of F in which the number of terms was finite but random, and the resulting random variables turned out to have the distribution F *exactly*.]

3 A Semi-Parametric Model and Algorithm for Random Effects Meta-Analysis

As an alternative to the class of mixtures of Dirichlet processes in Model (2.1), we may use a class of mixtures of conditional Dirichlet processes, in which μ is the median of F with probability one. This can be done by using a construction given in Doss (1985), which is reviewed below. Conditional Dirichlets have also been used to deal with identifiability issues by Newton et al. (1996).

3.1 A Model Based on Mixtures of Conditional Dirichlet Processes

Let α be a finite measure on the real line, and let $\mu \in (-\infty, \infty)$ be fixed. Let α_-^μ and α_+^μ be the restrictions of α to $(-\infty, \mu)$ and (μ, ∞), respectively, in the following sense. For any set A,

$$\alpha_-^\mu(A) = \alpha\{A \cap (-\infty, \mu)\} + \tfrac{1}{2}\alpha\{A \cap \{\mu\}\} \text{ and } \alpha_+^\mu(A) = \alpha\{A \cap (\mu, \infty)\} + \tfrac{1}{2}\alpha\{A \cap \{\mu\}\}.$$

Choose $F_- \sim \mathcal{D}_{\alpha_-^\mu}$ and $F_+ \sim \mathcal{D}_{\alpha_+^\mu}$ independently, and form F by

$$F(t) = \tfrac{1}{2}F_-(t) + \tfrac{1}{2}F_+(t). \tag{3.1}$$

The distribution of F will be denoted $\mathcal{D}_\alpha^\mu$. Note that with probability one, the median of F is μ. The prior $\mathcal{D}_\alpha^\mu$ has the following interpretation: If $F \sim \mathcal{D}_\alpha$, then $\mathcal{D}_\alpha^\mu$ is the conditional distribution of F given that the median of F is μ. Note that if α has median μ, then $E(F(t)) = \alpha(t)/\alpha\{(-\infty, \infty)\}$, as before, and furthermore, the quantity $\alpha\{(-\infty, \infty)\}$ continues to play the role of a precision parameter.

We will use a conditional Dirichlet process instead of a Dirichlet process; thus, Model (2.1) is replaced by the following:

$$\begin{aligned}
&\text{Conditional on } \psi_i, && D_i \overset{\text{ind}}{\sim} \mathcal{N}(\psi_i, \sigma_i^2), \quad i = 1, \ldots, m && \text{(3.2a)}\\
&\text{Conditional on } F, && \psi_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, m && \text{(3.2b)}\\
&\text{Conditional on } \mu, \tau, && F \sim \mathcal{D}^\mu_{M\mathcal{N}(\mu,\tau^2)} && \text{(3.2c)}\\
&\text{Conditional on } \tau, && \mu \sim \mathcal{N}(c, d\tau^2) && \text{(3.2d)}\\
& && \gamma = 1/\tau^2 \sim \text{Gamma}(a, b) && \text{(3.2e)}
\end{aligned}$$

If we wish to have a dispersed prior on (μ, τ) we may take a and b small, $c = 0$, and d large.

Before proceeding, we remark on the objectives when using a model of the form (3.2). Obtaining good estimates of the entire mixture distribution F usually requires very large sample sizes (here the number of studies). It will often be the case that the number of studies is not particularly large, and when using Model (3.2) the focus will then be on inference concerning the univariate parameter μ. For the case where the ψ_i's are observed completely (equivalent to the model where each study involves an infinite sample size), a model of the form (3.2) reduces to the model studied in Doss (1985), which was introduced for the purpose of robust estimation of μ. For this case, the posterior distribution of μ is given in Proposition 1 below. Assuming that the ψ_i's are all distinct, this posterior has a density that is essentially a product of two terms, one which shrinks it towards the mean of the ψ_i's, and the other which shrinks it towards their median. The mean of this posterior has good small sample (i.e. small m) frequentist properties (Doss 1983, 1985). This is discussed in more detail in Section 3.2 below, where we also discuss the ramifications for Model (3.2), where the ψ_i's are not necessarily all distinct. The general effect is illustrated in Section 4.1, where we compare Models (2.1) and (3.2).

To conclude, there are two reasons for using this model. First, as mentioned earlier, in (3.2) the parameter μ has a well-defined role, and as will be seen in Section 3.2 below, is easily estimated because it emerges as part of the Gibbs sampler output. Second, for Model (3.2) the posterior distribution of μ is not heavily influenced by a few outlying studies.

3.2 A Gibbs Sampler for Estimating the Posterior Distribution

There is no known way to obtain the posterior distribution in closed form for Model (3.2), and one must use Markov chain Monte Carlo. Markov chain methods for estimating the posterior distribution of F given D_i, $i = 1, \ldots, m$ in a model of the sort (2.1) are now well established, and are based on Escobar's (1994) use of the Pólya urn scheme of Blackwell and MacQueen (1973). It is possible to improve Escobar's (1994) original algorithm to substantially speed up convergence; see the recent paper by Neal (2000), which reviews previous work and presents new ideas. Here, we describe a basic Gibbs sampling algorithm for Model (3.2). It is possible

to develop versions of some of the algorithms described in Neal (2000) that can be implemented for Model (3.2), using as basis the formulas developed in this section, but we do not do so here.

We are primarily interested in the posterior distribution of μ, but we will also be interested in other posterior distributions, such as $\mathcal{L}_{\boldsymbol{D}}(\psi_{m+1})$ and $\mathcal{L}_{\boldsymbol{D}}(F)$. Here, ψ_{m+1} denotes the study-specific effect for a future study, so that $\mathcal{L}_{\boldsymbol{D}}(\psi_{m+1})$ is a predictive distribution. All of these can be estimated if we can generate a sample from $\mathcal{L}_{\boldsymbol{D}}(\boldsymbol{\psi}, \mu, \tau)$. Our Gibbs sampler on $(\boldsymbol{\psi}, \mu, \tau)$ has cycle $m+1$ and proceeds by updating $\psi_1, \ldots, \psi_m$ and then the *pair* (μ, τ).

One of the steps of our Gibbs sampling algorithm requires the following result, which gives the posterior distribution of the mixing parameter for the situation in which the ψ_i's are known. Let H be a distribution function. Define

$$H_\theta(x) = H\big((x-\mu)/\tau\big) \quad \text{for } \theta = (\mu, \tau).$$

Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)$. We will use $\boldsymbol{\psi}_{(-i)}$ to denote $(\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_m)$. (For the sake of completeness, we give the result for the general case where M_θ depends on θ.)

Proposition 1 *Assume that H is absolutely continuous, with continuous density h, and that the median of H is 0. If $\psi_1, \ldots, \psi_m$ are $\overset{\text{iid}}{\sim} F$, and if the prior on F is the mixture of conditional Dirichlets $\int \mathcal{D}^{\mu}_{M_\theta H_\theta}\, \lambda(d\theta)$, then the posterior distribution of θ given $\psi_1, \ldots, \psi_m$ is absolutely continuous with respect to λ and is given by*

$$\lambda_{\boldsymbol{\psi}}(d\theta) = c(\boldsymbol{\psi}) \left(\prod\nolimits^{\text{dist}} h\Big(\frac{\psi_i - \mu}{\tau}\Big) \right) K(\boldsymbol{\psi}, \theta) \left[\frac{(M_\theta)^{\#(\boldsymbol{\psi})} \Gamma(M_\theta)}{\Gamma(M_\theta + n)} \right] \lambda(d\theta), \tag{3.3}$$

where

$$K(\boldsymbol{\psi}, \theta) = \Big[\Gamma\big(M_\theta/2 + \textstyle\sum_{i=1}^m I(\psi_i < \mu)\big)\, \Gamma\big(M_\theta/2 + \sum_{i=1}^m I(\psi_i > \mu)\big) \Big]^{-1}, \tag{3.4}$$

the 'dist' in the product indicates that the product is taken over distinct values only, $\#(\boldsymbol{\psi})$ is the number of distinct values in the vector $\boldsymbol{\psi}$, Γ is the gamma function, and $c(\boldsymbol{\psi})$ is a normalizing constant.

Proposition 1 is proved through a computation much like the one used to prove Theorem 1 of Doss (1985). We need to proceed with that calculation with the linear Borel set A replaced by the product of Borel sets $A_1 \times A_2$, where A_1 is a subset of the reals and A_2 is a subset of the strictly positive reals, and use the fact that these rectangles form a determining class.

Note that if M_θ does not depend on θ, in (3.3) the term in square brackets is a constant that can be absorbed into the overall normalizing constant, and $K(\boldsymbol{\psi}, \theta)$ depends on θ only through μ (by slight abuse of notation we will then write $K(\boldsymbol{\psi}, \mu)$). In this case, (3.3) is similar to the familiar formula that says that the posterior is proportional to the likelihood times the prior, except that the likelihood is based on the distinct observations only, and we now also have the multiplicative factor $K(\boldsymbol{\psi}, \mu)$. If the prior on F was the mixture of Dirichlets $\int \mathcal{D}_{M_\theta H_\theta}\, \lambda(d\theta)$—as opposed to a mixture of conditional Dirichlets—the posterior distribution of θ would be the same as (3.3), but without the factor $K(\boldsymbol{\psi}, \theta)$ [Lemma 1 of Antoniak (1974)].

The factor $K(\boldsymbol{\psi}, \mu)$ plays an interesting role. Viewed as a function of μ, $K(\boldsymbol{\psi}, \mu)$ has a maximum when μ is at the median of the ψ_i's, and as μ moves away from the sample median in either direction, it is constant between the observations, and decreases by jumps at each observation. It has the general effect of shrinking the posterior distribution of μ towards the sample median, and does so more strongly when M is small.

Consider the term $L(\boldsymbol{\psi}, \theta) = \prod^{\text{dist}} h((\psi_i - \mu)/\tau)$ in (3.3). When implementing the Gibbs sampler, the effect of $L(\boldsymbol{\psi}, \theta)$ diminishes as M decreases. This is because when M is small, the vector $\boldsymbol{\psi}$ is partitioned into a batch of clusters, with the ψ_i's in the same cluster being equal. As a consequence, for small M, inference on μ is stable in Model (3.2) [because of the presence of the term $K(\boldsymbol{\psi}, \theta)$] but not in Model (2.1). It is an interesting fact that for small M the decreased relevance of the term $L(\boldsymbol{\psi}, \theta)$ creates problems for the estimation of μ, but *not* for the estimation of $\eta = \int x\, dF(x)$. We see this as follows. The conditional distribution of F given $\psi_1, \ldots, \psi_m$ is equal to

$$\mathcal{L}_{\boldsymbol{\psi}}(F) = \int \mathcal{L}_{\boldsymbol{\psi}}(F \mid \theta)\, \lambda_{\boldsymbol{\psi}}(d\theta) = \int \mathcal{D}_{MH_\theta + \sum_{i=1}^m \delta_{\psi_i}}\, \lambda_{\boldsymbol{\psi}}(d\theta),$$

and for small M, this is essentially equal to $\mathcal{D}_{\sum_{i=1}^m \delta_{\psi_i}}$, i.e. the mixing parameter plays no role.

Before giving the detailed description of the algorithm, we give an outline of the main steps involved.

Step 1: Update $\boldsymbol{\psi}$. For $i = 1, \ldots, m$, we generate successively ψ_i given the current values of ψ_j, $j \neq i$, μ, τ, and the data. The conditional distribution involved is a mixture of a normal truncated to the interval $(-\infty, \mu)$, another normal truncated to the interval (μ, ∞) and point masses at the ψ_j, $j \neq i$.

Step 2: Update (μ, τ). To generate (μ, τ) given $\boldsymbol{\psi}$ we go through two steps.

- 2a: We generate μ from its marginal distribution given $\boldsymbol{\psi}$ (τ being integrated out). This is proportional to a t-distribution times an easily calculated factor.
- 2b: We generate τ from its conditional distribution given μ and $\boldsymbol{\psi}$. The distribution of $1/\tau^2$ is a gamma.

We now describe the algorithm in detail, and we first discuss $\mathcal{L}_{\boldsymbol{D}}(\psi_i \mid \boldsymbol{\psi}_{(-i)}, \mu, \tau)$, where $\boldsymbol{\psi}_{(-i)} = (\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_m)$. Rewriting this as

$$\mathcal{L}_{\{\boldsymbol{\psi}_{(-i)}, \mu, \tau\}}(\psi_i \mid \boldsymbol{D}) = \mathcal{L}_{\{\boldsymbol{\psi}_{(-i)}, \mu, \tau\}}(\psi_i \mid D_i)$$

makes it simple to see that we can calculate this using a standard formula for the posterior distribution (i.e. the posterior is proportional to the likelihood times the prior). Using the well-known fact that

$$\text{if } X_1, \ldots, X_m \text{ are } \overset{\text{iid}}{\sim} F,\ F \sim \mathcal{D}_{MH}, \text{ then } \mathcal{L}(X_i \mid \boldsymbol{X}_{(-i)}) = \frac{MH + \sum_{j \neq i} \delta_{X_j}}{M + m - 1},$$

it is not too difficult to see that the "prior" is

$$\mathcal{L}_{\{\boldsymbol{\psi}_{(-i)}, \mu, \tau\}}(\psi_i) = \frac{1}{2} \frac{M\mathcal{N}_-^\mu(\mu, \tau^2) + \sum_{j \neq i;\ \psi_j < \mu} \delta_{\psi_j}}{M/2 + m_-} + \frac{1}{2} \frac{M\mathcal{N}_+^\mu(\mu, \tau^2) + \sum_{j \neq i;\ \psi_j > \mu} \delta_{\psi_j}}{M/2 + m_+}, \quad (3.5)$$

where

$$m_- = \sum_{j \neq i} I(\psi_j < \mu) \quad \text{and} \quad m_+ = \sum_{j \neq i} I(\psi_j > \mu),$$

and we are using the notation $\mathcal{N}^\mu_-(a,b)$ and $\mathcal{N}^\mu_+(a,b)$ to denote the restrictions (without renormalization) of the $\mathcal{N}(a,b)$ distribution to $(-\infty,\mu)$ and (μ,∞), respectively. The "likelihood" is

$$\mathcal{L}_{\{\boldsymbol{\psi}_{(-i)},\mu,\tau\}}(D_i \mid \psi_i) = \mathcal{L}(D_i \mid \psi_i) = \mathcal{N}(\psi_i, \sigma_i^2), \tag{3.6}$$

where the first equality in (3.6) follows because $\boldsymbol{\psi}_{(-i)}$, μ, and τ affect D_i only through their effect on ψ_i, and the second equality in (3.6) is just the model statement (2.1a).

To combine (3.5) and (3.6), we note that the densities of the subdistribution functions $\mathcal{N}^\mu_-(\mu,\tau^2)$ and $\mathcal{N}^\mu_+(\mu,\tau^2)$ are just the density of the $\mathcal{N}(\mu,\tau^2)$ distribution multiplied by the indicators of the sets $(-\infty,\mu)$ and (μ,∞), respectively. When we multiply these by the density of the $\mathcal{N}(\psi_i,\sigma_i^2)$ distribution, we can complete the square, resulting in constants times the density of a new normal distribution. This gives

$$\mathcal{L}_{\boldsymbol{D}}(\psi_i \mid \boldsymbol{\psi}_{(-i)},\mu,\tau) \;\propto\; C_-\mathcal{N}^\mu_-(A,B^2) + C_+\mathcal{N}^\mu_+(A,B^2) + \frac{\sum_{\substack{j\neq i\\ \psi_j<\mu}} \delta_{\psi_j} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big[-\frac{(D_i-\psi_j)^2}{2\sigma_i^2}\Big]}{M/2+m_-}$$
$$+ \frac{\sum_{\substack{j\neq i\\ \psi_j>\mu}} \delta_{\psi_j} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big[-\frac{(D_i-\psi_j)^2}{2\sigma_i^2}\Big]}{M/2+m_+}, \tag{3.7}$$

where

$$A = \frac{\mu\sigma_i^2 + D_i\tau^2}{\sigma_i^2+\tau^2}, \quad B^2 = \frac{\sigma_i^2\tau^2}{\sigma_i^2+\tau^2},$$

$$C_- = \frac{M/\big(\frac{M}{2}+m_-\big)}{\sqrt{2\pi(\sigma_i^2+\tau^2)}} \exp\Big[-\frac{(D_i-\mu)^2}{2(\sigma_i^2+\tau^2)}\Big] \quad \text{and} \quad C_+ = \frac{M/\big(\frac{M}{2}+m_+\big)}{\sqrt{2\pi(\sigma_i^2+\tau^2)}} \exp\Big[-\frac{(D_i-\mu)^2}{2(\sigma_i^2+\tau^2)}\Big].$$

The normalizing constant in (3.7) is

$$C_-\Phi\Big(\frac{\mu-A}{B}\Big) + C_+\Big(1-\Phi\Big(\frac{\mu-A}{B}\Big)\Big) + \frac{\sum_{\substack{j\neq i\\ \psi_j<\mu}} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big[-\frac{(D_i-\psi_j)^2}{2\sigma_i^2}\Big]}{M/2+m_-} +$$
$$\frac{\sum_{\substack{j\neq i\\ \psi_j>\mu}} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big[-\frac{(D_i-\psi_j)^2}{2\sigma_i^2}\Big]}{M/2+m_+},$$

(Φ is the standard normal cumulative distribution function) and this enables us to sample from $\mathcal{L}_{\boldsymbol{D}}(\psi_i \mid \boldsymbol{\psi}_{(-i)},\mu,\tau)$.

To generate (μ,τ) from $\mathcal{L}_{\boldsymbol{D}}(\mu,\tau \mid \boldsymbol{\psi})$, note that $\mathcal{L}_{\boldsymbol{D}}(\mu,\tau \mid \boldsymbol{\psi}) = \mathcal{L}(\mu,\tau \mid \boldsymbol{\psi})$, and this last is given by Proposition 1. Because of the factor $K(\boldsymbol{\psi},\mu)$, (3.3) is not available in closed form. However, we are taking λ to be given by (3.2d) and (3.2e), which is conjugate to the $\mathcal{N}(\mu,\tau^2)$ family, and this simplifies the form of (3.3). It is possible to generate (μ,τ) from (3.3) if we first generate μ from its marginal posterior distribution (this is proportional

to a t distribution multiplied by the factor $K(\boldsymbol{\psi}, \mu)$) and then generate τ from its conditional posterior distribution given μ ($1/\tau^2$ is just a Gamma).

In more detail, let m^* be the number of distinct ψ_i's, and $\bar{\psi}^* = (\sum^{\text{dist}} \psi_i)/m^*$, where the 'dist' in the sum indicates that the sum is taken over distinct values only. Then $\mathcal{L}(\mu, \tau \mid \boldsymbol{\psi})$ has density proportional to the product

$$g_{\boldsymbol{\psi}}(\mu, \tau) K(\boldsymbol{\psi}, \mu), \tag{3.8}$$

where $g_{\boldsymbol{\psi}}(\mu, \tau)$ has the form (3.2d)–(3.2e), with updated parameters a', b', c', d' given by

$$a' = a + m^*/2, \qquad b' = b + \frac{1}{2} \sum^{\text{dist}} (\psi_i - \bar{\psi}^*)^2 + \frac{m^*(\bar{\psi}^* - c)^2}{2(1 + m^* d)}$$

$$c' = \frac{c + m^* d \bar{\psi}^*}{m^* d + 1}, \qquad d' = \frac{1}{m^* + d^{-1}}$$

This follows from the conjugacy (Berger 1985, p. 288). Integrating out τ in (3.8) (and noting that $K(\boldsymbol{\psi}, \mu)$ does not depend on τ) we see that the (marginal) conditional distribution of μ given $\boldsymbol{\psi}$ has density proportional to

$$t\big(2a',\, c',\, b'd'/a'\big)(\cdot)\, K(\boldsymbol{\psi}, \cdot). \tag{3.9}$$

Here, $t(d, l, s^2)$ denotes the density of the t distribution with d degrees of freedom, location l, and scale parameter s. Because $K(\boldsymbol{\psi}, \cdot)$ is a step function, it is possible to do exact random variable generation from this posterior density. [The real line is partitioned into the $m^* + 1$ disjoint intervals formed by the m^* distinct ψ_i's. Then, the measures of these $m^* + 1$ intervals are calculated under (3.9) and renormalized to form a probability vector. One of these $m^* + 1$ intervals is chosen according to this probability vector, and finally a random variable is generated from the $t\big(2a',\, c',\, b'd'/a'\big)$ distribution restricted to the interval and renormalized to be a probability measure.] The conditional distribution of $1/\tau^2$ given $\boldsymbol{\psi}$ and μ is Gamma$\big(a' + 1/2,\, b' + (\mu - c')^2/2d'\big)$.

The algorithm described above gives us a sequence $\big(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)}\big)$, $g = 1, \ldots, G$, approximately distributed according to $\mathcal{L}_{\boldsymbol{D}}(\boldsymbol{\psi}, \mu, \tau)$, and as mentioned earlier, various posterior distributions can be estimated from this sequence. For example, to estimate $\mathcal{L}_{\boldsymbol{D}}(\psi_{m+1})$, we express this quantity as $\int \mathcal{L}_{\boldsymbol{D}}(\psi_{m+1} \mid \boldsymbol{\psi}, \mu, \tau)\, d\mathcal{L}_{\boldsymbol{D}}(\boldsymbol{\psi}, \mu, \tau)$, which we estimate by an average of G distributions each of the form (3.5).

Shrinkage in the Posterior Distribution If F is chosen from a Dirichlet process prior and $\psi_1, \ldots, \psi_m$ are $\overset{\text{iid}}{\sim} F$, then there will be ties among the ψ_i's, i.e. they will form clusters, and this tendency to form clusters is stronger when M is smaller. This fact is well known, and can easily be seen from Sethuraman's (1994) construction, for example. We now discuss the impact of this property on the posterior distribution of $\psi_1, \ldots, \psi_m$. In a standard parametric hierarchical model, i.e. (3.2) except that (3.2b) and (3.2c) are replaced with the simpler $\psi_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2)$, $i = 1, \ldots, m$, the posterior distribution of ψ_i involves shrinkage towards D_i and towards a grand mean. If we consider (3.7), we see that in the semi-parametric model, the posterior distribution of ψ_i is also shrunk towards those ψ_j's that are close to D_i. When M is small, the constants C_- and C_+ are small, and this means that this effect is stronger.

To conclude, the posterior distribution of ψ_i is affected by the results of all studies, but is more heavily affected by studies whose results are similar to those of study i. This effect is illustrated in the example of Section 4.1.

3.3 Connection Between Posterior Distributions Under Models (2.1) and (3.2)

It is natural to ask what is the connection between the posterior distribution of μ under our Model (3.2) and its posterior when the prior on F is (2.1). One way to get some insight into this question is as follows. Let $\nu^c(\boldsymbol{\psi}, \mu, \tau)$ denote the distribution of $(\boldsymbol{\psi}, \mu, \tau)$ under (3.2) for some specification of $(M, \{H_\theta\}, \lambda)$, and let $\nu^u(\boldsymbol{\psi}, \mu, \tau)$ denote the distribution under (2.1), for the same specification of these hyperparameters. Let $\nu^c_{\boldsymbol{D}}(\boldsymbol{\psi}, \mu, \tau)$ and $\nu^u_{\boldsymbol{D}}(\boldsymbol{\psi}, \mu, \tau)$ denote the corresponding posterior distributions. Proposition 2 below gives the relationship between these two posterior distributions.

Proposition 2 *Assume that H is absolutely continuous, with continuous density h, and that the median of H is 0. Then the Radon-Nikodym derivative $[d\nu^c_{\boldsymbol{D}}/d\nu^u_{\boldsymbol{D}}]$ is given by*

$$\left[\frac{d\nu^c_{\boldsymbol{D}}}{d\nu^u_{\boldsymbol{D}}}\right](\boldsymbol{\psi}, \mu, \tau) = AK(\boldsymbol{\psi}, \mu),$$

where K is given in (3.4), and A does not depend on $(\boldsymbol{\psi}, \mu, \tau)$.

The proof is given in the Appendix. Here we remark on how the proposition can be used. In general, A is difficult to compute, but this need not be a problem. Suppose that $\left(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)}\right)$, $g = 1, \ldots, G$ is Markov chain output generated under Model (2.1). Expectations with respect to Model (3.2) can be estimated by using a weighted average of the $\left(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)}\right)$'s, where the vector $\left(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)}\right)$ is given weight proportional to $K\left(\boldsymbol{\psi}^{(g)}, \mu^{(g)}\right)$ (Hastings 1970), and to do this we do not need to know A. Thus, vectors $\left(\boldsymbol{\psi}^{(g)}, \mu^{(g)}, \tau^{(g)}\right)$ such that $\mu^{(g)}$ is far from the median of $\psi_1^{(g)}, \ldots, \psi_m^{(g)}$ are given lower weight than vectors for which $\mu^{(g)}$ is close to the median of $\psi_1^{(g)}, \ldots, \psi_m^{(g)}$. Burr et al. (2003) applied this reweighting scheme on the output of a simpler program that runs the Markov chain for Model (2.1), in order to arrive at their estimates.

A drawback of the reweighting approach is that when the two distributions differ greatly, a few of the Markov chain points will take up most of the weight, and the result will be that estimates are unstable unless the chain is run for an extremely large number of cycles.

4 Illustrations

Here we illustrate the use of our models on two meta-analyses. In the first example, the issue of main interest is the basic question of whether or not there is evidence of a treatment effect, and the focus is on the latent parameter μ. In the second example, the principal interest is on the latent parameters ψ_i's. In each case we ran our Gibbs sampler for $100{,}000$ cycles and discarded the first $5{,}000$.

4.1 Decontamination of the Digestive Tract

Infections acquired in intensive care units are an important cause of mortality. One strategy for dealing with this problem involves selective decontamination of the digestive tract. This is designed to prevent infection by preventing carriage of potentially pathogenic micro-organisms from the oropharynx, stomach, and gut. A meta-analysis of 22 randomized trials to investigate the benefits of selective decontamination of the digestive tract was carried out by an international collaborative group (Selective Decontamination of the Digestive Tract Trialists' Collaborative Group 1993 [henceforth DTTCG 1993]). In each trial, patients in an intensive care unit were randomized to either a treatment or a control group. The treatments varied, with some including a topical (non-absorbable) antibiotic, while others included in addition a systemic antibiotic. The antibiotics varied across trials. In each trial, the proportion of individuals who acquired an infection was recorded for the treatment and control groups and an odds ratio was reported. The authors of the paper used a fixed effects model, in which the 22 trials were assumed to measure the same quantity. The 22 odds ratios were combined via the Mantel-Haenszel-Peto method. The results were that there is overwhelming evidence that selective decontamination is effective in reducing the risk of infection: a 95% confidence interval for the common odds ratio was found to be $(.31, .43)$. As expected, owing to the large variation in treatment across trials, a test of heterogeneity was significant (p-value $< .001$); however, a frequentist random effects analysis (DerSimonian and Laird 1986) gave similar results.

This data set was reconsidered by Smith et al. (1995), who used a Bayesian hierarchical model in which for each trial there is a true log odds ratio ψ_i, viewed as a latent variable, and for which the observed log odds ratio is an estimate. The true log odds ratios are assumed to come from a normal distribution with mean μ and variance τ^2, and a prior is put on (μ, τ). Smith et al. (1995) show that the posterior probability that μ is negative is extremely close to 1, confirming the results of DTTCG (1993), although with a different model.

An interesting later section of DTTCG (1993) considers mortality as the outcome variable. Each of the 22 trials reported also the proportion of individuals who died for the treatment and control groups, and again an odds ratio was reported. Using a fixed effects model, DTTCG (1993) find that the results are much less clear cut. The common odds ratio was estimated to be $.90$, with a 95% confidence interval of $(.79, 1.04)$. In 14 of the studies the treatment included both topical and systemic antibiotics, and medical considerations suggest that the effect of the treatment would be stronger in these studies. Indeed, for this subgroup the common odds ratio was estimated to be $.80$, with a 95% confidence interval of $(.67, .97)$, which does not include 1. (Consideration of this subgroup had been planned prior to the analysis of the data.) The data for these 14 studies appear in lines 2–5 of Table 1, and the odds ratios appear in line 6 of the table. The studies are arranged in order of increasing odds ratio, rather than in the original order given in DTTCG (1993), to facilitate inspection of the table.

Smith et al. (1995) noted that (for infection as the outcome variable), the data do not seem to follow a normal distribution, and for this reason also used a t-distribution. It is difficult to test for normality for this kind of data (the studies had different sizes, and there are only 14 studies). Nevertheless, the normal probability plot given in Figure 1 suggests some deviation from normality for our data as well. In meta-analysis studies it is often the case that there is some kind of grouping in the data—studies that have similar designs may yield similar results.

Some evidence of this appears in Figure 1.

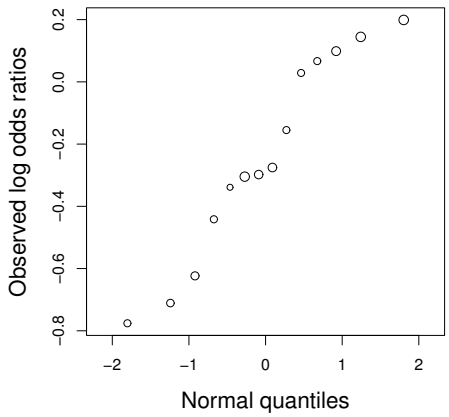

Figure 1: Normal probability plot for the mortality data for studies including both topical and systemic antibiotics. The areas of the circles are proportional to the inverse of the standard error.

We analyze these data using our semi-parametric Bayesian model. We fit model (3.2) with $a = b = .1$, $c = 0$, and $d = 1000$, and several values of M including $.1$ and 1000, in order to assess the effect of this hyperparameter on the conclusions. Large values of M essentially corresponds to a parametric Bayesian model based on a normal distribution, and this asymptopia is for practical purposes reached for M as little as 20. The value $M = .1$ gives an extreme case, and would not ordinarily be used (see Sethuraman and Tiwari 1982); it is included here only to give some insight into the behavior of the model.

Figure 2 below gives the posterior distribution of μ for $M = 1000$ and 1. For $M = 1000$, the posterior probability that μ is positive is $.04$, but for $M = 1$, the posterior probability is the non-significant figure of $.16$, which suggests that while the treatment reduces the rate of infection, there is not enough evidence to conclude that there is a corresponding reduction in mortality, even if the treatment involves both topical and systemic antibiotics.

The middle group of lines in Table 1 give the means of the posterior distributions for five values of M. The average of the posterior means (over the 14 studies) is very close to $.84$ for each of the five values of M. From the table we see that for $M = 1000$, the posterior means of the odds ratios are all shrunk from the value Obs OR towards $.84$. But for small values of M, the posterior means are also shrunk towards the average for the observations in their vicinity.

To get some understanding of the difference in behavior between Models (3.2) and (2.1) we ran a Gibbs sampler appropriate for Model (2.1), from which we created the last three lines of Table 1. From the table we make the following observations. First, the two models give virtually identical results for large M, as one would expect. Second, the shrinkage in Model (2.1) seems to have a different character, with attraction towards an overall mean being much stronger for small values of M. To get some insight on how the two models handle outliers, we did the following experiment. We took the study with the largest observed odds ratio and changed that from 1.2 to 2.0 (keeping that observation's standard error and everything else the same), and reran the algorithms. For large M, for either model $E(\psi_{14} \mid \boldsymbol{D})$ moved

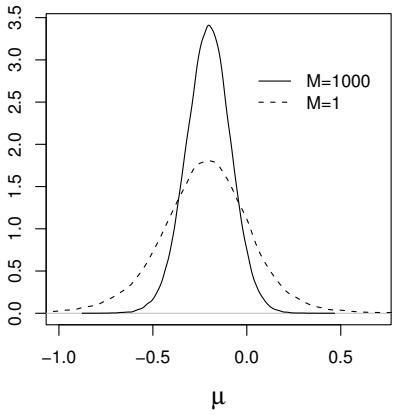

Figure 2: Posterior distribution of μ for $M = 1000$ and 1.

from 1.03 to 1.50. For $M = 1$, for Model (3.2) $E(\psi_{14} \mid \boldsymbol{D})$ moved from 1.04 to 1.44, whereas for Model (2.1) $E(\psi_{14} \mid \boldsymbol{D})$ moved from $.93$ to 1.52, a more significant change.

4.2 Clopidogrel vs. Aspirin Trial

When someone suffers an atherosclerotic vascular event (such as a stroke or a heart attack), it is standard to administer an antiplatelet drug, to reduce the chance of another event. The oldest such drug is Aspirin, but there are many other drugs on the market. One of these is Clopidogrel, and an important study (CAPRIE Steering Committee 1996) compared this drug with Aspirin in a large-scale trial. In this study, $19{,}185$ individuals who had suffered a recent atherosclerotic vascular event were randomized to receive either Clopidogrel or Aspirin. Patients were recruited over a three-year period, and mean follow-up was 1.9 years. There were 939 events over $17{,}636$ person-years for the Clopidogrel group, compared with 1021 events over 17,519 person-years for the Aspirin group, giving a risk ratio for Clopidogrel vs. Aspirin of $.913$ [with 95% confidence interval $(.835, 0.997)$], and Clopidogrel was judged superior to Aspirin, with a two-sided p-value of $.043$. As a consequence of this landmark study, Clopidogrel was favored over the much cheaper Aspirin for patients who have had an atherosclerotic vascular event.

A short section of the paper (p. 1334) discusses briefly the outcomes for three subgroups of patients, those participating in the study because they had had a stroke, a heart attack [myocardial infarction (MI)], or peripheral arterial disease (PAD). The results for the three groups differed: the risk ratios for Clopidogrel vs. Aspirin were 0.927, 1.037, and 0.762 for the stroke, MI, and PAD groups, respectively. A test of homogeneity of the three risk ratios gives a p-value of $.042$, providing mild evidence against the hypothesis that the three risk ratios are equal.

We will analyze the data using the methods developed in this paper. On the surface, this data set does not fit the description in Section 1 of this paper. The study was indeed designed as a multicenter trial involving 384 centers. However, the protocol was so well defined that

Study no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Treat inf:	14	22	27	11	4	51	33	24	14	14	15	34	45	47
Treat tot:	45	55	74	75	28	131	91	161	49	48	51	162	220	220
Cont inf:	23	33	40	16	12	65	40	32	15	14	14	31	40	40
Cont tot:	46	57	77	75	60	140	92	170	47	49	50	160	220	220
Obs OR	0.46	0.49	0.54	0.64	0.71	0.74	0.74	0.76	0.86	1.03	1.07	1.10	1.16	1.22
$M = .1$	0.67	0.66	0.66	0.76	0.82	0.78	0.80	0.81	0.87	0.93	0.95	1.02	1.05	1.06
$M = 1$	0.70	0.69	0.69	0.78	0.83	0.79	0.80	0.81	0.87	0.92	0.93	0.99	1.02	1.04
$M = 5$	0.72	0.71	0.71	0.78	0.83	0.79	0.80	0.81	0.85	0.90	0.91	0.97	1.00	1.03
$M = 20$	0.72	0.71	0.71	0.78	0.83	0.79	0.80	0.81	0.85	0.90	0.91	0.96	1.00	1.03
$M = 1000$	0.72	0.71	0.71	0.78	0.83	0.79	0.80	0.81	0.85	0.90	0.91	0.96	1.01	1.04
Model (2.1)														
$M = 1$	0.79	0.78	0.78	0.82	0.84	0.82	0.83	0.83	0.85	0.86	0.87	0.89	0.91	0.93
$M = 5$	0.74	0.73	0.73	0.80	0.83	0.80	0.81	0.82	0.85	0.89	0.89	0.94	0.98	1.00
$M = 20$	0.72	0.72	0.72	0.79	0.83	0.79	0.80	0.81	0.85	0.89	0.90	0.96	1.00	1.03

Table 1: Odds ratios for 14 studies. Lines 2 and 3 give the number infected and the total number in the treatment group, respectively, and lines 4 and 5 give the same information for the control group. The line "Obs OR" gives the odds ratios that are observed in the 14 studies. The next five lines give the means of the posteriors under the semi-parametric Model (3.2) for five values of M, and the bottom three lines give posterior means for Model (2.1).

it is reasonable to ignore the center effect. We focus instead on the patient subgroups. Of course, one may object to using a random effects model when there are only three groups involved. However, there is nothing in our Bayesian formulation that requires us to have a large number of groups involved—we simply should not expect to be able to obtain accurate estimates of the overall parameter μ and especially of the mixing distribution F. We carry out this analysis for two reasons. First, an analysis by subgroup is of medical interest, in that for the MI subgroup, Aspirin seems to outperform Clopidogrel, or at the very least, the evidence in favor of Clopidogrel is weaker. Second, there are studies currently under way that are comparing the two drugs for patients with several sets of gene profiles. Thus the kind of analysis we do here will apply directly to those studies, and the shrinkage discussed earlier may give useful information. Before proceeding we remark that in this situation there is no particular reason for preferring Model (3.2) to (2.1) (the emphasis is on the ψ_i's) and in any case the two models turn out to give virtually identical conclusions.

Let $\psi_{\text{stroke}}, \psi_{\text{MI}}$, and ψ_{PAD} be the logs of the true risk ratios for the three groups. Their corresponding estimates are $(D_{\text{stroke}}, D_{\text{MI}}, D_{\text{PAD}}) = (-0.076, 0.036, -0.272)$. We work on the log scale because the normal approximation in (3.2a) is more accurate on this scale. The estimated standard errors of the D's are 0.067, 0.083, and 0.091, respectively. (These are obtained from the number of events and the number of patient-years given on p. 1334 of the paper, using standard formulas.) We fit model (3.2) with a, b, c and d as in Section 4.1. The posterior distributions of the ψ's, on the original scale, are shown in Figure 3 for $M = 1$ and 1000, the latter essentially corresponding to a parametric Bayesian model. As expected, there is shrinkage towards an overall value of $.91$. The shrinkage is stronger for smaller values of M.

Our conclusions are as follows. The superiority of Clopidogrel to Aspirin for PAD is

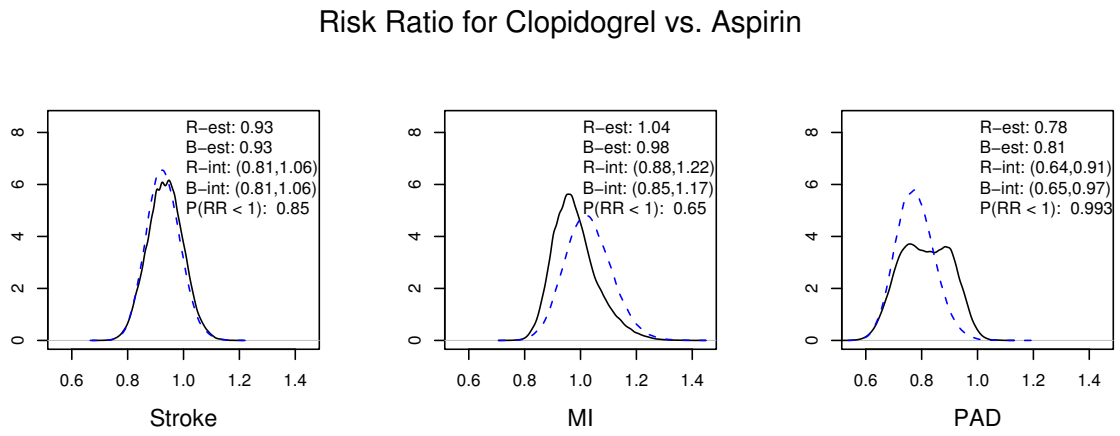

Figure 3: Posterior distributions of risk ratios, for three subgroups for Model (3.2) using $M = 1$ (solid lines) and $M = 1000$ (dashed lines). In each plot, R-est is the estimate reported in the Caprie paper; B-est is the mean of the posterior distribution; R-int is the 95% confidence interval reported in the paper; B-int is the central 95% probability interval for the posterior distribution; P(RR < 1) is the posterior probability that the risk ratio is less than one. All posteriors are calculated for the case $M = 1$, although the corresponding quantities for $M = 1000$ are virtually identical.

unquestionable, no matter what analysis is done. For stroke, the situation is less clear: the posterior probability that Clopidogrel is better is about $.85$ (for a wide range of values of M.) For MI, in the Bayesian model, the posterior probability that Clopidogrel is superior to Aspirin is $.38$ for $M = 1000$. Even after the stronger shrinkage that occurs for $M = 1$, this probability is only $.65$, so there is insufficient evidence for preferring Clopidogrel. The Caprie report states that "the administration of Clopidogrel to patients with atherosclerotic vascular disease is more effective than Aspirin ..." Our analysis suggests that this recommendation does not have an adequate basis for the MI subgroup.

Implementation of the Gibbs sampling algorithm of Section 3.2 is done through easy to use S-PLUS/R functions, available from the authors upon request. All the simulations can be done in S-PLUS/R, although considerable gain in speed can be obtained by calling dynamically loaded C subroutines.

Appendix: Proof of Proposition 2

A very brief outline of the proof is as follows. To calculate the Radon-Nikodym derivative $[d\nu^c_{\boldsymbol{D}}/d\nu^u_{\boldsymbol{D}}]$ at the point $(\boldsymbol{\psi}^{(0)}, \mu^{(0)}, \tau^{(0)})$, we will find the ratio of the probabilities, under $\nu^c_{\boldsymbol{D}}$ and $\nu^u_{\boldsymbol{D}}$, of the $(m+2)$-dimensional cubes centered at $(\boldsymbol{\psi}^{(0)}, \mu^{(0)}, \tau^{(0)})$ and of width ϵ, let ϵ tend to 0, and give a justification for why this gives the Radon-Nikodym derivative.

We now proceed with the calculation, which is not entirely trivial. Let $\theta_0 = (\mu_0, \tau_0) \in \Theta$ and $\boldsymbol{\psi}^{(0)} = (\psi^{(0)}_1, \ldots, \psi^{(0)}_m) \in \mathbb{R}^m$ be fixed. Let $\psi^{(0)}_{(1)} < \cdots < \psi^{(0)}_{(r)}$ be the distinct values of $\psi^{(0)}_1, \ldots, \psi^{(0)}_m$, and let $m_1, \ldots, m_r$ be their multiplicities (we will write $\psi_{(j)}$ instead of $\psi^{(0)}_{(j)}$ and

μ instead of μ_0 to lighten the notation, whenever this will not cause confusion). Let

$$m_- = \sum_{i=1}^{m} I(\psi_i < \mu), \quad m_+ = \sum_{i=1}^{m} I(\psi_i > \mu),$$

and

$$r_- = \sum_{j=1}^{r} I(\psi_{(j)} < \mu), \quad r_+ = \sum_{j=1}^{r} I(\psi_{(j)} > \mu).$$

For small $\epsilon > 0$, let $C^{\epsilon}_{\psi_i^{(0)}} = (\psi_i^{(0)} - \epsilon/2, \psi_i^{(0)} + \epsilon/2)$, and define $C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}$ to be the cube $C^{\epsilon}_{\psi_1^{(0)}} \times \cdots \times C^{\epsilon}_{\psi_m^{(0)}}$. Similarly define $B^{\epsilon}_{\theta_0}$ in Θ space. To calculate the likelihood ratio we first consider the probability of the set $\{\theta \in B^{\epsilon}_{\theta_0},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}$ under the two measures ν^c and ν^u (and not $\nu^c_{\boldsymbol{D}}$ and $\nu^u_{\boldsymbol{D}}$). We will find it convenient to associate probability measures with their cumulative distribution functions, and to use the same symbol to refer to both. Denoting the set of all probability measures on the real line by $\mathcal{P}$, we have

$$\begin{aligned}
\frac{\nu^c\{\theta \in B^{\epsilon}_{\theta_0},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}}{\nu^u\{\theta \in B^{\epsilon}_{\theta_0},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}} &= \frac{\nu^c\{\theta \in B^{\epsilon}_{\theta_0},\, F \in \mathcal{P},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}}{\nu^u\{\theta \in B^{\epsilon}_{\theta_0},\, F \in \mathcal{P},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}} \\
&= \frac{\displaystyle\int_{B^{\epsilon}_{\theta_0}} \int_{\mathcal{P}} \prod_{i=1}^{m} F\{C^{\epsilon}_{\psi_i^{(0)}}\}\, \mathcal{D}^{\mu}_{\alpha_\theta}(dF)\, \lambda(d\theta)}{\displaystyle\int_{B^{\epsilon}_{\theta_0}} \int_{\mathcal{P}} \prod_{i=1}^{m} F\{C^{\epsilon}_{\psi_i^{(0)}}\}\, \mathcal{D}_{\alpha_\theta}(dF)\, \lambda(d\theta)}. \qquad \text{(A.1)}
\end{aligned}$$

Let

$$g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{\substack{1\le i\le m \\ \psi_i<\mu}} F\{C^{\epsilon}_{\psi_i^{(0)}}\}\, \mathcal{D}^{\mu}_{\alpha_\theta}(dF), \quad g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{\substack{1\le i\le m \\ \psi_i>\mu}} F\{C^{\epsilon}_{\psi_i^{(0)}}\}\, \mathcal{D}^{\mu}_{\alpha_\theta}(dF), \quad \text{(A.2)}$$

and

$$g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{1\le i\le m} F\{C^{\epsilon}_{\psi_i^{(0)}}\}\, \mathcal{D}_{\alpha_\theta}(dF). \qquad \text{(A.3)}$$

Take ϵ small enough so that the sets $(\psi^{(0)}_{(j)} - \epsilon/2, \psi^{(0)}_{(j)} + \epsilon/2)$, $j = 1, \ldots, r$ are disjoint. Using the definition of F given in (3.1), including the independence of F_- and F_+, we rewrite the inner integral in the numerator of (A.1) and obtain

$$\begin{aligned}
\frac{\nu^c\{\theta \in B^{\epsilon}_{\theta_0},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}}{\nu^u\{\theta \in B^{\epsilon}_{\theta_0},\, \boldsymbol{\psi} \in C^{\epsilon}_{\boldsymbol{\psi}^{(0)}}\}} &= \frac{\displaystyle\int_{B^{\epsilon}_{\theta_0}} g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)\, g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)\, \lambda(d\theta)}{\displaystyle\int_{B^{\epsilon}_{\theta_0}} g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)\, \lambda(d\theta)} \\
&= \frac{\displaystyle\int_{B^{\epsilon}_{\theta_0}} \left(\frac{g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r_-} \prod_{\substack{1\le i\le m \\ \psi_i<\mu}} (m_i-1)!}\right)\left(\frac{g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r_+} \prod_{\substack{1\le i\le m \\ \psi_i>\mu}} (m_i-1)!}\right) \lambda(d\theta)}{\displaystyle\int_{B^{\epsilon}_{\theta_0}} \left(\frac{g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r} \prod_{1\le i\le m} (m_i-1)!}\right) \lambda(d\theta)}. \qquad \text{(A.4)}
\end{aligned}$$

We may rewrite $g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)$ and $g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)$ in (A.2) as

$$g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{\substack{1\le j\le r_- \\ \psi_{(j)}<\mu}} \Big[F\big\{C^{\epsilon}_{\psi^{(0)}_{(j)}}\big\}\Big]^{m_j} \mathcal{D}^{\mu}_{\alpha_\theta}(dF), \tag{A.5a}$$

and

$$g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{\substack{r_-+1\le j\le r \\ \psi_{(j)}>\mu}} \Big[F\big\{C^{\epsilon}_{\psi^{(0)}_{(j)}}\big\}\Big]^{m_j} \mathcal{D}^{\mu}_{\alpha_\theta}(dF), \tag{A.5b}$$

respectively, and rewrite $g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)$ in (A.3) as

$$g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \int_{\mathcal{P}} \prod_{1\le j\le r} \Big[F\big\{C^{\epsilon}_{\psi^{(0)}_{(j)}}\big\}\Big]^{m_j} \mathcal{D}_{\alpha_\theta}(dF). \tag{A.6}$$

(Note that the conditions $\psi_{(j)} < \mu$ and $\psi_{(j)} > \mu$ in the products in (A.5a) and (A.5b) are redundant, since the $\psi_{(j)}$'s are ordered.) Let $A_j(\epsilon) = \alpha_\theta\big\{(\psi_{(j)} - \epsilon/2, \psi_{(j)} + \epsilon/2)\big\}$, $j = 1,\dots,r_-$, and also define $A_{r_-+1}(\epsilon) = (M/2) - \sum_{j=1}^{r_-} A_j(\epsilon)$. Calculation of (A.5a) is routine since it involves only the finite-dimensional Dirichlet distribution. The integral in (A.5a) is $E(U_1^{m_1}\cdots U_{r_-}^{m_{r_-}})$ where $(U_1,\dots,U_{r_-},U_{r_-+1}) \sim \mathrm{Dirichlet}\big(A_1(\epsilon),\dots,A_{r_-+1}(\epsilon)\big)$, and we can calculate this expectation explicitly. We obtain

$$g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta) = \Big(\frac{1}{2}\Big)^{m_-} \frac{\Gamma\big(\frac{M}{2}\big)}{\Big(\prod_{j=1}^{r_-}\Gamma\big(A_j(\epsilon)\big)\Big)\Gamma\big(A_{r_-+1}(\epsilon)\big)} \, \frac{\Big(\prod_{j=1}^{r_-}\Gamma\big(A_j(\epsilon)+m_j\big)\Big)\Gamma\big(A_{r_-+1}(\epsilon)\big)}{\Gamma\big(\frac{M}{2}+m_-\big)}.$$

Let

$$f^{-}_{\boldsymbol{\psi}^{(0)}}(\theta) = \Big(\frac{1}{2}\Big)^{m_-}\Big(\prod_{j=1}^{r_-} h_\theta(\psi_{(j)})\Big)\frac{M^{r_-}\,\Gamma\big(\frac{M}{2}\big)}{\Gamma\big(\frac{M}{2}+m_-\big)},$$

$$f^{+}_{\boldsymbol{\psi}^{(0)}}(\theta) = \Big(\frac{1}{2}\Big)^{m_-}\Big(\prod_{j=r_-+1}^{r} h_\theta(\psi_{(j)})\Big)\frac{M^{r_+}\,\Gamma\big(\frac{M}{2}\big)}{\Gamma\big(\frac{M}{2}+m_+\big)},$$

and

$$f_{\boldsymbol{\psi}^{(0)}}(\theta) = \Big(\prod_{j=1}^{r} h_\theta(\psi_{(j)})\Big)\frac{M^{r}\,\Gamma(M)}{\Gamma(M+m)}.$$

Here, h_θ is the density of H_θ. Using the recursion $\Gamma(x+1) = x\Gamma(x)$ and the definition of the derivative, we see that

$$\frac{g^{-,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r_-}\prod_{j=1}^{r_-}(m_j-1)!} \to f^{-}_{\boldsymbol{\psi}^{(0)}}(\theta) \qquad \text{for each } \theta\in\Theta.$$

Similarly,

$$\frac{g^{+,\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r_+}\prod_{j=r_-+1}^{r}(m_j-1)!} \to f^{+}_{\boldsymbol{\psi}^{(0)}}(\theta) \quad \text{and} \quad \frac{g^{\epsilon}_{\boldsymbol{\psi}^{(0)}}(\theta)}{\epsilon^{r}\prod_{j=1}^{r}(m_j-1)!} \to f_{\boldsymbol{\psi}^{(0)}}(\theta) \qquad \text{for each } \theta\in\Theta.$$

Furthermore, since h is continuous, the convergence is uniform in small neighborhoods of θ_0. Therefore, it is clear that (A.4) converges to

$$\left(\frac{1}{2}\right)^m \frac{\Gamma^2\left(\frac{M}{2}\right)\Gamma(M+m)}{\Gamma(M)\Gamma\left(\frac{M}{2}+m_-\right)\Gamma\left(\frac{M}{2}+m_+\right)} = \left(\frac{1}{2}\right)^m \left[\frac{\Gamma^2\left(\frac{M}{2}\right)\Gamma(M+m)}{\Gamma(M)}\right] K(\boldsymbol{\psi}^{(0)}, \mu^{(0)}), \quad \text{(A.7)}$$

and we note that the expression in brackets does not depend on $(\boldsymbol{\psi}^{(0)}, \mu^{(0)}, \tau^{(0)})$.

We now return to $\nu^c_{\boldsymbol{D}}$ and $\nu^u_{\boldsymbol{D}}$. The posterior is proportional to the likelihood times the prior, and since the likelihood is the same under the two models (i.e. (2.1a) and (3.2a) are identical), we obtain

$$\left[\frac{d\nu^c_{\boldsymbol{D}}}{d\nu^u_{\boldsymbol{D}}}\right](\boldsymbol{\psi}^{(0)}, \mu^{(0)}, \tau^{(0)}) = AK(\boldsymbol{\psi}^{(0)}, \mu^{(0)}),$$

where A does not depend on $(\boldsymbol{\psi}^{(0)}, \mu^{(0)}, \tau^{(0)})$, as stated in the proposition.

The calculation of the Radon-Nikodym derivatives $[d\nu^c/d\nu^u]$ and $[d\nu^c_{\boldsymbol{D}}/d\nu^u_{\boldsymbol{D}}]$ in this way is supported by a martingale construction (see e.g. Durrett 1991, pp. 209–210) and the main theorem in Pfanzagl (1979).

References

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis (Second Edition)*, Springer-Verlag.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Burr, D., Doss, H., Cooke, G., and Goldschmidt-Clermont, P. (2003), "A Meta-Analysis of Studies on the Association of the Platelet PlA Polymorphism of Glycoprotein IIIa and Risk of Coronary Heart Disease," *Statistics in Medicine*, 22, 1741–1760.

CAPRIE Steering Committee (1996), "A Randomised, Blinded, Trial of Clopidogrel versus Aspirin in Patients at Risk of Ischaemic Events (CAPRIE)," *Lancet*, 348, 1329–1339.

Carlin, J. B. (1992), "Meta-Analysis for 2×2 Tables: A Bayesian Approach," *Statistics in Medicine*, 11, 141–158.

DerSimonian, R. and Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177–188.

Dey, D., Müller, P., and Sinha, D. (1998), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer-Verlag Inc.

Doss, H. (1983), "Bayesian Nonparametric Estimation of Location," Ph.D. thesis, Stanford University.

— (1985), "Bayesian Nonparametric Estimation of the Median: Part I: Computation of the Estimates," *The Annals of Statistics*, 13, 1432–1444.

— (1994), "Bayesian Nonparametric Estimation for Incomplete Data Via Successive Substitution Sampling," *The Annals of Statistics*, 22, 1763–1786.

DuMouchel, W. (1990), "Bayesian Metaanalysis," in *Statistical Methodology in the Pharmaceutical Sciences*, Marcel Dekker (New York), pp. 509–529.

Durrett, R. (1991), *Probability: Theory and Examples*, Brooks/Cole Publishing Co.

Escobar, M. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," Ph.D. thesis, Yale University.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Gelfand, A. E. and Kottas, A. (2002), "A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289–305.

Ghosh, M. and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal (Disc: P76-93)," *Statistical Science*, 9, 55–76.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Morris, C. N. and Normand, S. L. (1992), "Hierarchical Models for Combining Information and for Meta-Analyses (Disc: P335-344)," in *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, Clarendon Press (Oxford), pp. 321–335.

Muliere, P. and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson-Dirichlet Priors," *The Canadian Journal of Statistics*, 26, 283–297.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Newton, M. A., Czado, C., and Chappell, R. (1996), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 142–153.

Pfanzagl, J. (1979), "Conditional Distributions as Derivatives," *The Annals of Probability*, 7, 1046–1050.

Selective Decontamination of the Digestive Tract Trialists' Collaborative Group (1993), "Meta-Analysis of Randomised Controlled Trials of Selective Decontamination of the Digestive Tract," *British Medical Journal*, 307, 525–532.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Sethuraman, J. and Tiwari, R. C. (1982), "Convergence of Dirichlet Measures and the Interpretation of Their Parameter," in *Statistical Decision Theory and Related Topics III, in two volumes*, Academic (New York; London), vol. 2, pp. 305–315.

Skene, A. M. and Wakefield, J. C. (1990), "Hierarchical Models for Multicentre Binary Response Studies," *Statistics in Medicine*, 9, 919–929.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995), "Bayesian Approaches to Random-effects Meta-Analysis: A Comparative Study," *Statistics in Medicine*, 14, 2685–2699.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996), *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5, (version ii)*, MRC Biostatistics Unit (Cambridge).