Bayesian treatment comparison using parametric mixture priors computed from elicited histograms

Peter F Thall, Moreno Ursino, Véronique Baudouin, Corinne Alberti, Sarah Zohar

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(2) 404–418

DOI: 10.1177/0962280217726803
journals.sagepub.com/home/smm
SAGE

Bayesian treatment comparison using parametric mixture priors computed from elicited histograms

Peter F Thall,[1] Moreno Ursino,[2] Véronique Baudouin,[3] Corinne Alberti[4] and Sarah Zohar[2]

Abstract

A Bayesian methodology is proposed for constructing a parametric prior on two treatment effect parameters, based on graphical information elicited from a group of expert physicians. The motivating application is a 70-patient randomized trial to compare two treatments for idiopathic nephrotic syndrome in children. The methodology relies on histograms of the treatment parameters constructed manually by each physician, applying the method of Johnson et al. (2010). For each physician, a marginal prior for each treatment parameter characterized by location and precision hyperparameters is fit to the elicited histogram. A bivariate prior is obtained by averaging the marginals over a latent physician effect distribution. An overall prior is constructed as a mixture of the individual physicians' priors. A simulation study evaluating several versions of the methodology is presented. A framework is given for performing a sensitivity analysis of posterior inferences to prior location and precision and illustrated based on the idiopathic nephrotic syndrome trial.

Keywords

Bayesian inference, clinical trial, mixture model, pediatric medicine, prior elicitation, rare diseases

1 Introduction

A pervasive problem when comparing treatments based on randomized clinical trials in children, rare diseases, or important disease subgroups, is that the sample size often is too small to obtain a confirmatory conclusion using conventional statistical methods. Examples of subgroups include patients with a biomarker believed to interact with treatment, an age interval arising due to metabolic heterogeneity in children, or cancer patients who have relapsed after achieving remission with frontline therapy. In such settings, even a multi-institution trial may not obtain a sample size large enough to provide convincing comparative inferences.

Depending on the setting, the treatment parameter may be the probability of a binary response, the mean of a real-valued outcome, or mean survival time. As a toy example to illustrate the sort of settings we have in mind, suppose that one wishes to compare the response probabilities, θ_1 and θ_2, of two competing treatments. If a randomized trial of 160 patients gives 39 responses in 75 (52%) patients for treatment 1 and 54 responses in 85 (64%) patients for treatment 2, then a frequentist two-sided binomial test of the null hypothesis $\theta_1 = \theta_2$ has p value $= .14$, which conventionally is considered nonsignificant. From a Bayesian viewpoint, if one assumes independent $beta(.50, .50)$ priors for θ_1 and θ_2, then the posterior probability that treatment 2 provides at least a .15 improvement over treatment 1 is $\Pr(\theta_1 + .15 < \theta_2 \mid data) = .32$, and the posterior 95% credible intervals (CIs) are .41–.63 for θ_1 and .53–.73 for θ_2, which overlap substantially. Thus, while the data suggest that treatment 2 is

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA
[2]UMRS 1138, CRC, INSERM, University Paris 5, University Paris 6, France
[3]Department of Pediatric Nephrology, University hospital Robert Debré-APHP, France
[4]UMR 1123, INSERM, Hôpital Robert-Debré, APHP, University Paris 7, France

Corresponding author:
Peter F Thall, Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA.
Email: rex@mdanderson.org

superior, conventional comparative inferences are far from confirmatory. Assuming the above sample response rates of 52 and 64%, for example, hypothetical data 130/250 and 185/290 having these rates but based on much larger samples would give nonoverlapping 95% posterior CIs .46–.58 and .58–69. This suggests that, if one were planning a trial using nonoverlapping posterior 95% CIs as a criterion for posterior reliability, a total sample size of roughly 540 would be required. As explained above, we are motivated by settings where a sample this large simply is not practical. Because the ultimate goal of a randomized trial is to provide a convincing basis for practicing physicians to choose between treatments, in settings where the sample size is not large and it is unlikely that a trial will be repeated, one reasonable course of action is seek expert opinion as an additional source of information. This leads naturally to a Bayesian approach wherein expert opinion is elicited and formalized by an informative prior distribution on (θ_1, θ_2).

Our proposed methodology was motivated by the desire to analyze the results of a randomized trial of two treatments for idiopathic nephrotic syndrome, which is the most common kidney disease in children. About 90% of cases are sensitive to corticosteroids, and among them about 60% have dependence on corticosteroids sufficient to justify the addition of an immunosuppressive agent to reduce the frequency of relapses and side effects.[1,2] Cyclophosphamide is an immunosuppressive agent often used as first-line therapy; and in case of failure, the second-line therapy used most often is cyclosporine. For both treatments, duration of administration is limited by toxicity, which occurs in the bone marrow and gonads with cyclophosphamide and in the kidneys with cyclosporine. The risk of toxicity is the primary reason that cyclophosphamide is administered on a short, three-month basis.[3] Observational studies have shown that, if no toxicity occurs, cyclophosphamide may achieve remission at one year in 17–67% of patients. A new treatment, mycophenolate mofetil (MMF), may reduce the corticosteroid dependence and thus limit the need for cyclosporine. Observational studies have shown that continuous treatment with MMF provides remission in 42–75% of patients.[4,5] A key motivation is that MMF has been shown to be nonnephrotoxic and nongonadotoxic, so if it can be established that MMF has a response rate similar to that of cyclophosphamide, then MMF would be preferable due to its superior safety.

Motivated by these results, a randomized trial (NEPHROMYCY, NCT01092962) was conducted to compare the efficacy of cyclophosphamide (148 mg/kg during 12 weeks) versus MMF (1200 mg/m^2 during 18 months) in children with steroid-dependent nephrotic syndrome. The primary outcome was response, defined as relapse not occurring during the first 24 months of follow-up. The trial included 70 patients from 26 pediatric nephrology centers in France. Denoting the response probabilities by θ_1 for cyclophosphamide and θ_2 for MMF, a key posterior probability is $\pi^E_{1,2}(\epsilon) = \Pr(\theta_1 - \epsilon < \theta_2 \mid data)$, computed for small $\epsilon = 0.05$ or 0.10. A large value of $\pi^E_{1,2}(\epsilon)$ provides evidence that MMF is "ϵ-equivalent" to cyclophosphamide in terms of their response probabilities. This may motivate a practicing physician to use MMF rather than cyclophosphamide, since MMF is nontoxic.

Because idiopathic nephrotic syndrome is a rare disease, it was recognized at the start that the trial's sample size would be small, and that 70 children could be expected to be accrued in a realistic time period. A Bayesian analysis of the trial data was planned in the protocol. To obtain prior expert opinion before the trial was begun, each of 17 physicians experienced in treating this disease was asked to construct a histogram reflecting what they believed to be the distribution of the probability of response for each of MMF and cyclophosphamide, denoted by θ_1 and θ_2. This was done by applying the so-called "bins-and-chips" graphical method of Johnson et al.,[6] which we will describe in detail below, in Section 3.

The general issue that we address in this article is how histograms elicited in this way from a set of experts may be used to construct a parametric prior on (θ_1, θ_2) as a basis for a Bayesian analysis to compare the two parameters. Our proposed methodology for constructing a parametric prior is carried out in three stages. In the first stage, a parametric distribution for the marginal priors of the θ_j's, with location parameter μ and precision parameter γ, is specified. For each expert and each θ_j, this model is fit to the elicited histogram to obtain a marginal parametric prior. In the second stage, a bivariate expert-specific prior for (θ_1, θ_2) is constructed by averaging the product of each expert's two marginal parametric priors over a distribution for two correlated latent expert effects, one for μ and the other for γ. We consider two ways to formulate this latent effect distribution, either assuming homogeneity across experts or including expert-specific covariates, if they are available. In the third stage, an overall joint prior for (θ_1, θ_2) is constructed as a discrete mixture of all the experts' bivariate parametric priors. For this mixture prior, each expert's weight may be a covariate-based index of their experience, or an index of the agreement between the means of the elicited histograms and corresponding model-based parameter estimates obtained from the data. A third approach is simply to weight the experts equally. Once the overall prior has been established, it may be used as a basis for Bayesian analyses of (θ_1, θ_2).

To address the issues of how prior location and precision may affect posterior inferences, we provide a formal framework for conducting a prior-to-posterior sensitivity analysis. We do this by constructing an array of alternative priors, each obtained by specifying numerical values of two quantities, one that changes the prior's location $E(\theta_2-\theta_1)$ and the other that changes its precision. Posterior quantities used to compare θ_1 to θ_2 are computed for each alternative prior. This produces an array of posterior values, one for each combination of location shift and precision transformation, including one based on the untransformed prior. This set of prior-to-posterior quantities may be used as a basis for making a conclusion about the comparative effectiveness of the two treatments, in light of both the observed data and the elicited prior opinion.

Making inferences about medical treatments based on small- to moderate-sized clinical trials while assuming informative priors constructed from elicited expert opinion is inherently controversial. While the ability to formally incorporate expert opinion *a priori* may be considered a major benefit of taking a Bayesian approach, it must be done carefully. Use of an informative prior constructed from elicited expert opinion may be seen as introducing bias into posterior inferences. The problems of eliciting expert opinion, constructing priors from the elicited values, and performing Bayesian analyses on that basis have been addressed by numerous authors in many different settings. Many authors have discussed methods for prior elicitation,[6–12] establishing priors for Bayesian model-based clinical trials and medical applications,[13–15] graphical methods for prior elicitation,[6,16,17] and combining priors and expert opinion.[18,19] A review is given by O'Hagan et al.[20] Our methodology is related to the general development for combining expert priors given by Albert et al.,[21] who consider the somewhat different problem of using elicited probabilities and elicited quantiles to construct a prior. Their framework requires an additional parameter quantifying prior uncertainty to be elicited from each expert, which is not required by our method.

Our proposed methodology for constructing a parametric prior for (θ_1, θ_2) based on elicited histograms, and the process of solving numerically for hyperparameters, will be presented in Section 2. In Section 3, we describe how the graphical bins-and-chips method of Johnson et al.[6] was applied to elicit histograms for the response probabilities of MMF and cyclophosphamide for the NEPHROMYCY trial, and how beta distributions were fit to the histograms. In Section 4, a simulation study is presented that compares six different versions of our proposed method, obtained from the two ways to formulate the latent physician effect distribution and the three ways to weight the experts. Section 5 describes a formal method for performing sensitivity analyses of posterior inferences to prior location and informativeness, and this is illustrated by a simulated version of the NEPHROMYCY data set. We close with a brief discussion in Section 6.

2 Parametric models for the priors

2.1 Definition of treatment parameters

For the *i-th* subject in the data set to be analyzed, $i = 1, \cdots, n,$ denote treatment by $\tau_i,$ observed outcome by $Y_i,$ and covariates by $\boldsymbol{Z}_i = (Z_{i,1}, \cdots, Z_{i,q})$. Index treatments by $j = 1, 2,$ and denote $\theta_{j,i} = E(Y_i \mid \tau_i = j, \boldsymbol{Z}_i)$. As in Wahed and Thall,[22] we define the overall effect of treatment j as the mean over the sample of n_j subjects

$$\bar{\theta}_j = \int \theta_{j,i}(z) f_{\boldsymbol{Z}}(z) \mathrm{d}z = \frac{1}{n_j} \sum_{i=1}^{n_j} \theta_{j,i} \tag{1}$$

where $f_{\boldsymbol{Z}}$ denotes the patient covariate distribution. That is, for each treatment $j = 1, 2,$ we define $\bar{\theta}_j$ by averaging over the empirical distribution of subject covariates, with subjects weighted equally within treatment groups. If subject covariates are not available and subjects are assumed to be homogeneous, then $\theta_{j,1} = \cdots = \theta_{j,n_j} = \bar{\theta}_j$. Note that we have elaborated the notation by now denoting the two overall treatment parameters as $\bar{\theta}_1$ and $\bar{\theta}_2$, rather than θ_1 and θ_2, as done previously in Section 1.

2.2 Probability models for the physicians' marginal priors

Let $k = 1, \cdots, K$ index the expert physicians from whom the histograms for $\bar{\theta}_1$ and $\bar{\theta}_2$ are elicited. We formulate a marginal model for the k-th physician's prior on $\bar{\theta}_j$ by assuming a parametric distribution $p_{j,k}(\bar{\theta}_j \mid \mu_{j,k}, \gamma_{j,k}),$ where $\mu_{j,k}$ is a location parameter and $\gamma_{j,k} > 0$ is a precision parameter, for $j = 1, 2$. We formulate the marginal distributions in terms of location and precision parameters to facilitate analysis of the sensitivity of posterior inferences to prior bias and informativeness, which we will describe in Section 5 below. Association between $\bar{\theta}_1$ and $\bar{\theta}_2$ in each physician's joint prior is induced by assuming a bivariate prior for two latent physician effects (frailties),

one effect for location and a second effect for precision. The marginals of $[\bar{\theta}_1 \mid \mu_{1,k}, \gamma_{1,k}]$ and $[\bar{\theta}_2 \mid \mu_{2,k}, \gamma_{2,k}]$ are defined conditional on the k-th physician's frailties, and a bivariate prior then is obtained by averaging the product of these conditional marginals over the frailty distribution.

Models for the marginal priors of the $\theta_{j,k}$'s may be chosen for their tractability, since they will be fit to the elicited histograms. We require that they are parameterized in terms of location and precision parameters, μ and γ, to give the structure needed for conducting prior-to-posterior sensitivity analyses. For binary Y, where the $\theta_{j,k}$'s are probabilities, the beta distribution is a convenient, flexible family of parametric priors. Suppressing (j, k) temporarily, the beta pdf with mean μ and variance $\mu(1-\mu)/(1+\gamma)$ is given by

$$p(x \mid \mu, \gamma) = \frac{x^{\mu\gamma-1}(1-x)^{(1-\mu)\gamma-1}}{B(\mu\gamma, (1-\mu)\gamma)}, \quad 0 < x < 1$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(\cdot)$ denotes the gamma function. Thus, larger γ corresponds to greater precision. For real-valued Y, the normal distribution with mean μ and precision parameter $\gamma = 1/var(\theta)$ is a natural choice for the prior family. For Y an event time or other nonnegative-valued random variable, there are several reasonable two-parameter models that may be defined in terms of location and precision parameters. For example, a flexible model for the prior of the $\theta_{j,k}$'s is a gamma distribution with mean μ and precision γ, with pdf

$$p(x \mid \mu, \gamma) = \frac{(\mu\gamma)^{\mu^2\gamma} x^{\mu^2\gamma-1} e^{-\mu\gamma x}}{\Gamma(\mu^2\gamma)} \quad x > 0$$

To obtain priors for $\bar{\theta}_1$ and $\bar{\theta}_2$, the parametric models $p_{1,k}(\bar{\theta}_1 \mid \mu_{1,k}, \gamma_{1,k})$ and $p_{2,k}(\bar{\theta}_2 \mid \mu_{2,k}, \gamma_{2,k})$ are fit to the corresponding histograms elicited from the k-th physician, which yields numerical values of the four hyperparameters $\mu_{1,k}, \gamma_{1,k}, \mu_{2,k}, \gamma_{2,k}$, for each $k = 1, \cdots, K$. A numerical method for obtaining these fits is described below, in Section 3.2.

Since the two marginal prior distributions $p_{1,k}$ and $p_{2,k}$ both are obtained from the k-th physician, this implies that, *a priori*, $\bar{\theta}_1$ and $\bar{\theta}_2$ may be associated with each other for each physician. To formalize this idea, we propose two similar but different approaches for constructing a bivariate prior $p_k(\bar{\theta}_1, \bar{\theta}_2)$ from the marginal priors $p_{1,k}$ and $p_{2,k}$, for each $k = 1, \cdots, K$. Both approaches rely on bivariate latent physician effects (frailties), which are conceptual variables that are not observed. The frailties are used to induce prior within-physician correlation between $\bar{\theta}_1$ and $\bar{\theta}_2$, and thus obtain a bivariate prior on these parameters for each physician. The physician frailites are motivated by the idea that, given the two histograms and resulting beta priors obtained from the k-th physician expert, the pairs $(\mu_{1,k}, \gamma_{1,k})$ and $(\mu_{2,k}, \gamma_{2,k})$ must be associated with each other through the unobserved physician frailty, which in turn implies that $\bar{\theta}_1$ and $\bar{\theta}_2$ are associated with each other for each physician.

2.3 First method for computing prior hyperparameters

To implement Method 1 for establishing bivariate physician-specific priors on $(\bar{\theta}_1, \bar{\theta}_2)$, for each $k = 1, \cdots, K$, we first link $\mu_{j,k}$ and $\gamma_{j,k}$ to linear terms, each of which is the sum of a real-valued parameter and a latent physician effect. Let $\epsilon_k = (\epsilon_{k,\mu}, \epsilon_{k,\gamma}), k = 1, \cdots, K$, be independent and identically distributed (iid) pairs of real-valued latent physician effects, following a bivariate normal distribution

$$\epsilon_k \sim N(\mathbf{0}, \boldsymbol{\Sigma}) = N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{\epsilon,\mu} & \rho\sigma_{\epsilon,\mu}\sigma_{\epsilon,\gamma} \\ \rho\sigma_{\epsilon,\mu}\sigma_{\epsilon,\gamma} & \sigma^2_{\epsilon,\gamma} \end{bmatrix}\right) \tag{2}$$

Denote $\boldsymbol{\sigma} = (\sigma_{\epsilon,\mu}, \sigma_{\epsilon,\gamma}, \rho)$, and denote this bivariate normal by $p_\epsilon(x_\mu, x_\gamma \mid \boldsymbol{\sigma})$ for $(x_\mu, x_\gamma) \in R^2$. Let g_μ and g_γ denote appropriate link functions. If each $\bar{\theta}_j$ is a probability, then g_μ may be the logit, probit, or complementary log–log link. The identity link or log link may be used, respectively, if $\mu_{j,k}$ is real-valued or positive real-valued. Since $\gamma_{j,k} > 0$ in any case, g_γ may be the log link. In our motivating application, g_μ is the logit link and g_γ is the log link.

For Method 1, we assume that

$$\begin{aligned} g_\mu(\mu_{j,k}) &= \upsilon_{j,k,\mu} + \epsilon_{k,\mu} \\ g_\gamma(\gamma_{j,k}) &= \upsilon_{j,k,\gamma} + \epsilon_{k,\gamma} \end{aligned} \tag{3}$$

where the $\upsilon_{j,k,\mu}$'s are real-valued location parameters and the $\upsilon_{j,k,\gamma}$'s are real-valued precision parameters. The joint prior of $(\bar{\theta}_1, \theta_2)$ for the k-th physician is obtained by averaging over the bivariate physician effect distribution. Denoting $\upsilon_k = (\upsilon_{1,k,\mu}, \upsilon_{2,k,\mu}, \upsilon_{1,k,\gamma}, \upsilon_{2,k,\gamma})$, the joint prior is

$$p_k(\bar{\theta}_1, \bar{\theta}_2 | \upsilon_k, \boldsymbol{\sigma}) = \int_{R^2} \left\{ \prod_{j=1,2} p_{j,k}(\bar{\theta}_j \mid g_\mu^{-1}(\upsilon_{j,k,\mu} + x_\mu), g_\gamma^{-1}(\upsilon_{j,k,\gamma} + x_\gamma)) \right\} p_\epsilon(x_\mu, x_\gamma | \boldsymbol{\sigma})\, dx_\mu dx_\gamma$$

Under this parametric model, the hyperparameters υ_k are specific to p_k only, whereas the hyperparameters $\boldsymbol{\sigma}$ that characterize Σ in the bivariate normal distribution of the ϵ_k's appear in all K physician's priors. The frailty prior $p_\epsilon(x_\mu, x_\gamma \mid \boldsymbol{\sigma})$ induces correlation and shrinks $\bar{\theta}_1$ and $\bar{\theta}_2$ toward each other, with the degree of shrinkage determined by the assumed numerical values of the entries of $\boldsymbol{\sigma}$.

To implement our Method 1, when applying the elicitation method of Johnson et al.,[6] there is no elicited prior information on the parameters $\boldsymbol{\sigma} = (\rho, \sigma_{\epsilon,\mu}, \sigma_{\epsilon,\gamma})$ that characterize the variance–covariance matrix Σ of the latent effect distribution that we have introduced. Thus, numerical values of these three hyperparameters must be specified. It may be argued that introduction of $(\epsilon_{k,\mu}, \epsilon_{k,\gamma})$ is an unnecessary complication, and a more parsimonious approach would be to assume that the two beta priors for $\bar{\theta}_1$ and $\bar{\theta}_2$ are independent. Alternatively, it may be argued that, if the latent physician effects are included in the model, then values of ρ, $\sigma_{\epsilon,\mu}$, and $\sigma_{\epsilon,\gamma}$ also should be elicited, that is, that our assumed model requires a more elaborate elicitation procedure. Since the meanings of these second-order parameters to a physician are not entirely straightforward, however, it is not obvious how such an additional elicitation may be carried out. As shown in Section 2.4 below, our second method for computing hyperparameters does provide numerical values of $\boldsymbol{\sigma}$, essentially because it exploits the information in physician covariates. Thus, to complete the prior specification when implementing Method 1, we use the numerical values of $\boldsymbol{\sigma}$ obtained by Method 2. Still, since in theory, any value of $\rho \in (-1, 1)$ may be specified; and moreover in some applications, physician covariates may not be available. In Section 4, we will present an analysis of the sensitivity of posterior inferences to ρ when using Method 1.

2.4 Second method for computing prior hyperparameters

The second approach, Method 2, for constructing physician-specific priors on $(\bar{\theta}_1, \bar{\theta}_2)$ incorporates physician covariate vectors, $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$, if they are available. Thus, for Method 2, the model for $p_{j,k}(\bar{\theta}_j \mid \mu_{j,k}, \gamma_{j,k})$ is extended to include regression structure. This approach is appropriate if it is desired that the prior should reflect physician covariate effects on the $\mu_{j,k}$'s and $\gamma_{j,k}$'s. Method 2 also uses the values of $\mu_{1,k}, \gamma_{1,k}, \mu_{2,k}, \gamma_{2,k}$ obtained by fitting the parametric models $p_{1,k}(\bar{\theta}_1 \mid \mu_{1,k}, \gamma_{1,k})$ and $p_{2,k}(\bar{\theta}_2 \mid \mu_{2,k}, \gamma_{2,k})$ to the elicited histograms, but in a very different way than Method 1. For Method 2, the latent physician effects are as before, but for each k, we assume that

$$\begin{aligned} g_\mu(\mu_{j,k}) &= \upsilon_{j,\mu} + \beta_\mu \boldsymbol{X}_k + \epsilon_{k,\mu} + e_\mu \\ g_\gamma(\gamma_{j,k}) &= \upsilon_{j,\gamma} + \beta_\gamma \boldsymbol{X}_k + \epsilon_{k,\gamma} + e_\gamma \end{aligned} \tag{4}$$

where $\mathbf{e} = (e_\mu, e_\gamma) \sim N_2\left(\mathbf{0}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix}\right)$ are general error terms associated with location and scale that do not vary with (j, k). Denoting $\beta = (\beta_\mu, \beta_\gamma)$, with Method 2, we define the marginals of $\bar{\theta}_1$ and $\bar{\theta}_2$ conditional on both $\boldsymbol{X}_k$ and ϵ_k, as

$$p_{j,k}(\bar{\theta}_j \mid \upsilon_{j,\mu}, \upsilon_{j,\gamma}, \beta, \boldsymbol{X}_k, \epsilon_k)$$

for $j = 1, 2$. In this regression formulation, there now are four intercept parameters, $\upsilon = (\upsilon_{1,\mu}, \upsilon_{1,\gamma}, \upsilon_{2,\mu}, \upsilon_{2,\gamma})$, in the linear terms, and these are identical for all physicians, since allowing them to vary with k would render the model nonidentifiable. This is a key difference from the model (3) used with Method 1, where the intercept parameters vary with physician $k = 1, \cdots, K$. Thus, for Method 2, between-physician variability is accounted for by their covariates.

For Method 2, the available physician covariate information allows numerical values of $(\upsilon, \beta, \boldsymbol{\sigma})$ in the physician covariate regression model to be computed. A numerical value of σ_0 must be specified, however. To implement Method 2, we obtain these hyperparameter values by treating the location and dispersion parameters, obtained

from the elicited histograms, as pseudo outcomes and the hyperparameter vector $(\upsilon, \beta, \boldsymbol{\sigma})$ as pseudo parameters, fit regression model (4), and use the estimated pseudo parameters as the prior means of $(\upsilon, \beta, \boldsymbol{\sigma})$ in the marginal priors $\{p_{j,k}(\bar{\theta}_j \mid \boldsymbol{X}_k, \upsilon, \beta, \boldsymbol{\sigma}), j = 1, 2, k = 1, \cdots, K\}$. This fit may be done in several different ways, all of which give very similar numerical results. We did this by assuming independent N (0,100) pseudo priors for the elements of υ and β, and inverse gamma or uniform distributions for the elements of Σ. Additional details are given below, in Section 3.2. The posterior means obtained from fitting this nonlinear Bayesian regression model were used as the hyperparameters for the physician-specific marginal priors. Given these marginal priors for the k-th physician, the joint prior of $(\bar{\theta}_1, \bar{\theta}_2)$ is obtained, as in Method 1, by averaging over the bivariate physician effect distribution

$$p_k(\bar{\theta}_1, \bar{\theta}_2 | \upsilon, \beta, \boldsymbol{\sigma}, \boldsymbol{X}_k) = \int_{R^2} \left[\prod_{j=1,2} p_{j,k} \left\{ \bar{\theta}_j \mid g_\mu^{-1}(\upsilon_{j,\mu} + \beta_\mu \boldsymbol{X}_k + x_\mu), g_\gamma^{-1}(\upsilon_{j,\gamma} + \beta_\gamma \boldsymbol{X}_k + x_\mu) \right\} \right] p_\epsilon(x_\mu, x_\gamma | \boldsymbol{\sigma}) \, dx_\mu dx_\gamma$$

For Method 2, the K bivariate priors have identical hyperparameter vectors, $(\upsilon, \beta, \boldsymbol{\sigma})$, and the prior p_k is specific to the k-th physician only through the covariate vector $\boldsymbol{X}_k$.

2.5 Mixture priors

Given the K bivariate physician-specific parametric priors obtained by either Method 1 or 2, let $\boldsymbol{w} = (w_1, \cdots, w_K)$ denote physician weights that sum to 1. Using Method 1, the combined prior of $(\bar{\theta}_1, \bar{\theta}_2)$ is defined to be the mixture

$$p(\bar{\theta}_1, \bar{\theta}_2 \mid \upsilon_1, \cdots, \upsilon_K, \boldsymbol{\sigma}) = \sum_{k=1}^{K} w_k \, p_k(\bar{\theta}_1, \bar{\theta}_2 \mid \upsilon_k, \boldsymbol{\sigma}) \tag{5}$$

and with Method 2 the combined prior is

$$p(\bar{\theta}_1, \bar{\theta}_2 \mid \upsilon, \beta, \boldsymbol{\sigma}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_K) = \sum_{k=1}^{K} w_k \, p_k(\bar{\theta}_1, \bar{\theta}_2 \mid \upsilon, \beta, \boldsymbol{\sigma}, \boldsymbol{X}_k) \tag{6}$$

The physician weights $\boldsymbol{w}$ may be determined in several different ways, three of which are described here. To construct $\boldsymbol{w}$ empirically, denoting the treatment of patient i by τ_i, one may first fit a likelihood for $[Y_i \mid \tau_i, \boldsymbol{Z}_i]$ to the data $\mathcal{D}_n$, and denote the maximum likelihood estimates of the parameters for each (j, i) by $\hat{\theta}_{j,i}^{(like)}$, with $\hat{\boldsymbol{\theta}}^{(like)}$ the $2n$-vector of these estimates. Alternatively, estimates may be obtained as posterior means computed under a Bayesian model with noninformative pseudo priors. For each $j = 1, 2$, we quantify the mean of θ_j obtained from the elicited histogram of physician k by its empirical mean, which we denote by $\hat{\theta}_{j,k}^{(elicited)}$. For the k-th physician, the agreement between the two mean vectors $\hat{\boldsymbol{\theta}}_k^{(elicited)} = (\hat{\theta}_{1,k}^{(elicited)}, \hat{\theta}_{2,k}^{(elicited)})$ computed from that physician's elicited histograms and the $2n$ likelihood data-based estimated mean vectors $\hat{\boldsymbol{\theta}}^{(like)}$ can be quantified by the mean absolute deviation

$$\left\| \hat{\boldsymbol{\theta}}_k^{(elicited)} - \hat{\boldsymbol{\theta}}^{(like)} \right\| = \frac{1}{2n} \sum_{j=1}^{2} \sum_{i=1}^{n} \left| \hat{\theta}_{j,k}^{(elicited)} - \hat{\theta}_{j,i}^{(like)} \right|$$

Since smaller $\left\| \hat{\boldsymbol{\theta}}_k^{(elicited)} - \hat{\boldsymbol{\theta}}^{(like)} \right\|$ corresponds to closer agreement, we define the physician weights to be

$$w_k = \frac{\left\| \hat{\boldsymbol{\theta}}_k^{(elicited)} - \hat{\boldsymbol{\theta}}^{(like)}) \right\|^{-1}}{\sum_{r=1}^{K} \left\| \hat{\boldsymbol{\theta}}_r^{(elicited)} - \hat{\boldsymbol{\theta}}^{(like)}) \right\|^{-1}}$$

If physician covariates are available, then an alternative way to define the physician weights using the covariates is as follows. Without loss of generality, assume that each physician covariate is positive-valued and that larger $X_{k,l}$

corresponds to greater reliability of physician k, such as $X_{k,l}$ being years of experience. The weights then can be defined as

$$w_k = \frac{1}{q}\sum_{l=1}^{q}\frac{X_{k,l}}{\sum_{r=1}^{K} X_{r,l}}$$

A larger value of $X_{k,l}/\sum_{r=1}^{K} X_{r,l}$ corresponds to greater reliability of the opinion of physician k, relative to the other physicians, in terms of the l-th covariate. The average over $l = 1, \cdots, q$ treats the covariates as being equally important. With this weighting scheme, the physician-specific priors computed using Method 2 use each physician's covariates twice, once to obtain the prior p_k and a second time to obtain the weight w_k. A third alternative is simply to weight the physicians equally by setting all $w_k = 1/K$.

3 Bins and chips prior elicitation

When planning the NEPHROMYCY trial, each physician's prior on each θ_j was elicited by applying the bins and chips method of Johnson et al.,[6] as follows. The Bayesian approach first was explained to the group of physicians at a pretrial planning meeting. This included explanations of the primary efficacy outcome, and how *a priori* belief is combined with data, by application of Bayes' theorem, to compute a posterior distribution for making inferences about key parameters. An example of Bayesian thinking was presented, in which a physician who sees a patient reporting a pain in their chest may have the prior belief that the patient's actual illness has probability distribution Pr(anxiety) = .10, Pr(myocardiac infarction) = Pr(MI) = .30, and Pr(pneumonia) = .60. Then, after obtaining the results of a chest X-ray and electrocardiogram, this information changes the physician's belief so that the new probabilities are Pr(anxiety) = .05, Pr(MI) = 0, and Pr(pneumonia) = .95.

It was next explained that analysis of the trial data would require each of the physicians to provide their own prior on the response probabilities with each treatment (M or C). It was then explained that each physician would receive, by mail, an envelope containing a questionnaire (given in the Supplementary Material) that, once they had filled it out, would characterize their prior. The items of the questionnaire were then explained, including how to carry out the so-called "bins and chips" construction of each prior histogram. They were told that the envelope would contain 40 colored stickers, and that 20 stickers would be used to construct each prior. It was explained that, for each treatment response probability, each sticker represented probability .05, and that they should place 20 stickers into the discrete intervals printed on the questionnaire so that the resulting histogram would represent their belief about the distribution of the response probability for that treatment. The intervals used in the questionnaire were [0, .05], [.06, .10],. . ., [.91, .95], [.96, 1.00]. They were told to carry out this exercise for each of the two treatments, and mail back the competed questionnaire. During this explanation, a graphical illustration was provided, including several examples of what a competed histogram might look like. This illustration was the figure with colored chips given in Appendix C of Johnson et al.[6]

3.1 Computing marginal prior hyperparameters

In this section, we explain how one may perform the computations to obtain the parameters of each marginal parametric prior from the corresponding elicited histogram. The computation is carried out in two steps, which we describe for binary outcomes. In the first step, for each physician $k = 1, \ldots, K$, each histogram $j = 1, 2$ is matched with a beta distribution, $p_{j,k}$, having mean and precision parameters $(\mu_{j,k}, \gamma_{j,k})$. At the end of the elicitation process, for each expert $k = 1, \cdots, K$, and treatment $j = 1, 2$, the histogram gives the elicited prior probability $\theta_{j,k,r}^{(elicited)}$ of $P_{j,k}(l_r < \theta_{j,k} < u_r)$ for each of the subintervals, used in the elicitation, that partition the domain of the $\theta_{j,k}$'s. For probabilities, r indexes the 20 subintervals [0, .05], [.06, .10],. . ., [.96, 1.0]. In practice, some of the $\theta_{j,k,r}^{(elicited)}$ values may be 0, corresponding to an elicited prior that has support in a proper subset of [0, 1]. Denote the r-vector of elicited values corresponding to the intervals by $\boldsymbol{\theta}_{j,k}^{(elicited)}$. To solve for the two hyperparameters $(\mu_{j,k}, \gamma_{j,k})$ of the beta, we match the elicited prior probabilities with the model-based prior probabilities by minimizing

$$\sum_{r=1}^{20}\left\{P_{j,k}(l_r < \theta_{j,k} < u_r|\mu_{j,k}, \gamma_{j,k}) - \theta_{r,j,k}^{(elicited)}\right\}^2$$

This might be done by applying the Nelder-Mead algorithm. For real or positive real-valued $\bar{\theta}_j$'s, the graphical histogram elicitation method may be implemented by first determining a range $[L_\theta, U_\theta]$ for the parameters, and then partitioning this range into subintervals of equal width. One then proceeds as before, by asking each physician to place 20 stickers each having probability mass .05 into the intervals to construct a prior histogram for each of the $\bar{\theta}_j$'s.

In the second step, for Method 2, denoting $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, we treat the estimates $(\boldsymbol{\mu}, \boldsymbol{\gamma})$ obtained by the above minimization as pseudo outcomes in the regression model given by (4), and treat $(\upsilon_{1,\mu},\ \upsilon_{2,\mu},\ \upsilon_{1,\gamma},\ \upsilon_{2,\gamma},\ \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\gamma, \boldsymbol{\sigma}_{\epsilon,\mu}, \boldsymbol{\sigma}_{\epsilon,\gamma}, \boldsymbol{\rho})$ as pseudo parameters to be estimated. To fit the Bayesian regression model to estimate the hyperparameter means, described earlier, we assumed independent noninformative normal pseudo priors for the covariate effects, $\boldsymbol{\beta}_\mu \sim N(\mathbf{0}, \boldsymbol{\sigma}^2_{\beta_\mu} \boldsymbol{I})$ and $\boldsymbol{\beta}_\gamma \sim N(\mathbf{0}, \sigma^2_{\beta_\gamma} \boldsymbol{I})$ where $\boldsymbol{I}$ is the identity matrix, with both prior variances $\sigma^2_{\beta_\mu}$ and $\sigma^2_{\beta_\gamma}$ suitably large. Similarly, independent noninformative normal pseudo priors were assumed for $\upsilon_{j,\mu}$ and $\upsilon_{j,\gamma}$, denoted by $\upsilon_{j,\mu} \sim N(0, \sigma^2_\upsilon)$ and $\upsilon_{j,\gamma} \sim N(0, \sigma^2_\upsilon)$, with σ^2_υ one order of magnitude larger than $\sigma^2_{\beta_\mu}$ and $\sigma^2_{\beta_\gamma}$. Moreover, independent noninformative inverse gamma, denoted by $IG(1, 1)$, pseudo priors were assumed for $\sigma_{\epsilon,\mu}$, $\sigma_{\epsilon,\gamma}$, whereas a uniform distribution in the interval $(-1, 1)$ was assumed for ρ. Computations were carried out in R, using a rstan package, which is available as a Supplementary file. To approximate the double integral over R^2 to compute $p_k(\bar{\theta}_1, \bar{\theta}_2 | \boldsymbol{\upsilon}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{X}_k)$, we used Monte Carlo sampling.

To illustrate the marginal parametric priors obtained from the elicited histograms, Figure1 gives the elicited histograms and fitted beta priors for θ_1 and θ_2 for three of the 17 physicians who participated in the elicitation process in planning the NEPHROMYCY trial. Since some physicians used less than 20 stickers for some histograms, in these cases, we normalized the histogram to have total probability mass 1 before fitting the corresponding beta distribution. Plots of the elicited histograms and fitted beta distributions for all 17 physicians are given in Supplementary Figure S1.

Contour plots of the distributions of the estimates of (μ,γ) from the beta distributions fit to the elicited histograms of the 17 physicians are given in Figure 2 (left-hand side for cyclophosphamide and right-hand side for MMF). Histograms of the marginal distributions of the physician-specific estimates of μ (along the top) and γ (on the right side) are also given. The histograms for μ_1 and μ_2 show that, on average, the physicians believed MMF to have a higher response probability than cyclophosphamide, although there was substantial between-physician variability. The histograms for both the precision parameters γ_1 and γ_2 were highly disperse, but had remarkably similar shapes with most mass between 30 and 70.

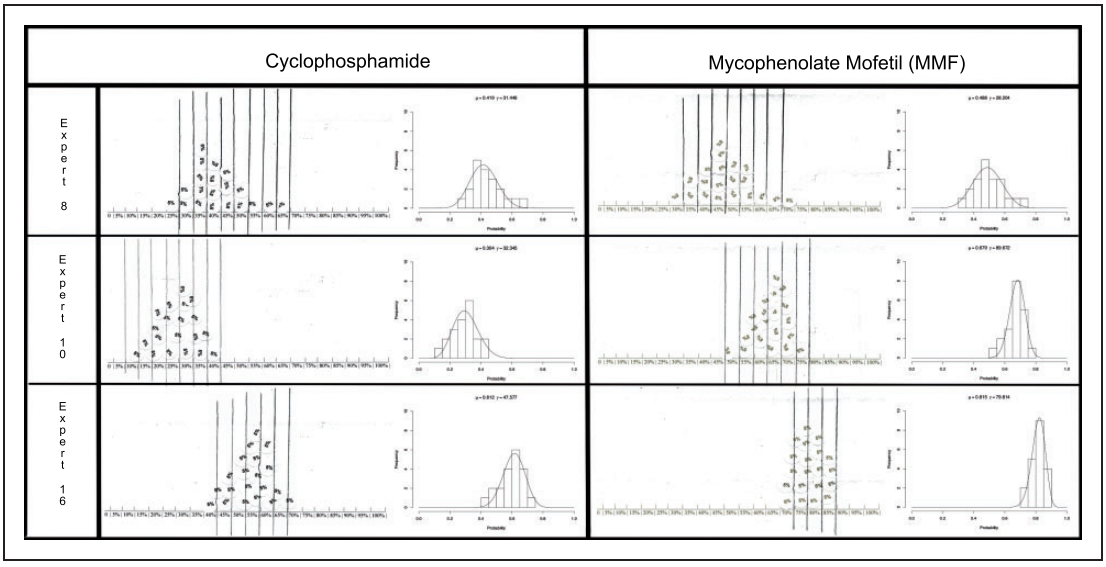

Figure 1. Elicited histograms and fitted beta priors for θ_1 (left-hand side – cyclophosphamide) and θ_2 (right-hand side – MMF) for three of the 17 physicians who participated in the elicitation process in planning the NEPHROMYCY trial.
MMF: mycophenolate mofetil.

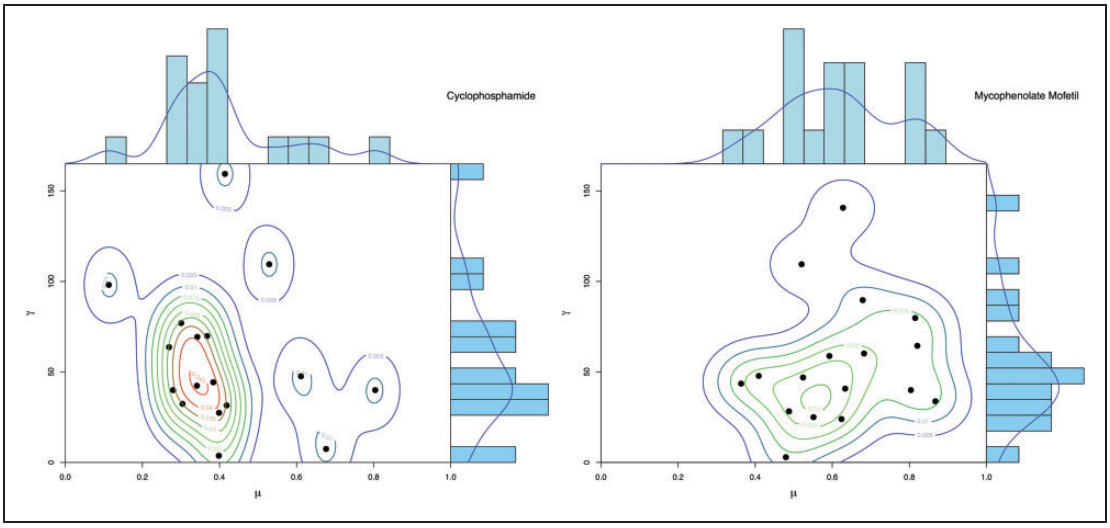

Figure 2. Contour plots of estimated (μ, γ) for each expert in the domain $(0, 1) \times R^+$ for the estimate prior response probabilities of cyclophosphamide (left-hand side) and MMF (right-hand side). Marginal histograms are plotted on the top for μ and right-hand side for γ.
MMF: mycophenolate mofetil.

4 Simulation study

In this section, we summarize a simulation study using Methods 1 and 2 and each of the three ways to weight physicians in the mixture prior, for a total of six versions of the methodology. Four scenarios were considered, determined by the assumed true values of the response probabilities $(p_1, p_2) = (0.5, 0.5)$, $(0.2, 0.3)$, $(0.2, 0.4)$, or $(0.4, 0.2)$. Let $\hat{\theta}_j^{est}$ denote the median of the marginal posterior density of θ_j, for $j = 1, 2$. Table 1 gives the posterior ϵ-equivalence probabilities $\pi_{12}^E(.05)$, $\pi_{12}^E(.10)$, and the median and first and third quantiles of $\hat{\boldsymbol{\theta}}_j^{est}$ obtained from 500 replications for each combination of Method and physician weighting scheme in each simulation scenario.

To implement Method 2 in the simulations, three physician covariates $\boldsymbol{X}_k$ were selected from the questionnaires given to the physicians when planning the NEPHROMYCY trial. These were the logarithm of the number of years experience as paediatrician, the logarithm of the average number of patients consulted per year, and a binary indicator of whether the physician had training in clinical trial methodology. These covariates were also used to compute the covariate-based physician weights in that version of the mixture prior.

The computed hyperparameters were

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.399 & -0.003 \\ -0.003 & 0.634 \end{bmatrix}$$

with $(\nu_{1,\mu}, \nu_{2,\mu}, \nu_{1,\gamma}, \nu_{2,\gamma}) = (-0.708,\ 0.237,\ 3.387,\ 3.395)$, $\boldsymbol{\beta}_\mu = (0.173,\ -0.049,\ -0.053)$, and $\boldsymbol{\beta}_\gamma = (-0.185,\ 0.239,\ 0.334)$. Figure 3 shows the two joint prior distributions for $(\bar{\theta}_1, \bar{\theta}_2)$ obtained using these values by Methods 1 and 2, using equal physician weights for the mixture. Method 2 produces a smoother surface, whereas Method 1 gives a bimodal distribution. This implies that that Method 2 is more informative than Method 1. As described earlier, the numerical $\boldsymbol{\Sigma}$ given above that was obtained via Method 2 will be assumed for both methods in what follows.

The simulation results are summarized in Table 1. In all scenarios and cases, there are at most trivial differences in the effects on posterior quantities of the three different ways to compute physician weights. For the null scenario where the true response probabilities in the two treatment groups are (0.5, 0.5), Method 2, which uses physician covariates to compute the prior, produces larger values, 0.96 to 0.97, of $\pi_{(1,2,\phi,\delta)}^E(.05)$, compared to the values 0.86 to 0.91 obtained using Method 1. This effect is seen for both $\pi_{(1,2,\phi,\delta)}^E(.05)$ and $\pi_{(1,2,\phi,\delta)}^E(.10)$ in the scenario with true response probabilities (0.4, 0.2), although in this case, the numerical values are far too small to provide convincing evidence of equivalence. The slightly greater dispersion produced by Method 2 is shown by the quartiles of the two posterior parameter distributions in all scenarios and cases.

Table 1. Simulations of a 70-subject trial using each combination of method for computing hyperparameters and weighting physicians.[a]

True $(\bar{\theta}_1, \bar{\theta}_2)$	Method	Physician weights	$\pi^E_{12}(0.05)$	$\pi^E_{12}(0.10)$	Posterior median (25th, 75th percentiles) $\bar{\theta}_1$	Posterior median (25th, 75th percentiles) $\bar{\theta}_2$
(0.5, 0.5)	1	1	0.86 (0.74, 0.95)	0.96 (0.9, 0.99)	0.48 (0.43, 0.53)	0.52 (0.47, 0.57)
		2	0.91 (0.79, 0.97)	0.97 (0.91, 0.99)	0.46 (0.41, 0.51)	0.54 (0.49, 0.59)
		3	0.88 (0.76, 0.96)	0.96 (0.9, 0.99)	0.47 (0.42, 0.52)	0.53 (0.48, 0.58)
	2	1	0.97 (0.89, 0.99)	0.99 (0.95, 1)	0.44 (0.4, 0.49)	0.56 (0.51, 0.61)
		2	0.97 (0.91, 0.99)	0.99 (0.96, 1)	0.44 (0.39, 0.49)	0.57 (0.52, 0.61)
		3	0.96 (0.9, 0.99)	0.99 (0.95, 1)	0.44 (0.4, 0.49)	0.56 (0.51, 0.61)
(0.2, 0.3)	1	1	0.98 (0.92, 1)	1 (0.98, 1)	0.21 (0.18, 0.24)	0.32 (0.27, 0.36)
		2	0.98 (0.93, 1)	1 (0.98, 1)	0.21 (0.18, 0.24)	0.34 (0.29, 0.39)
		3	0.98 (0.92, 0.99)	1 (0.98, 1)	0.21 (0.18, 0.25)	0.33 (0.28, 0.38)
	2	1	0.99 (0.97, 1)	1 (0.99, 1)	0.20 (0.17, 0.23)	0.36 (0.31, 0.4)
		2	1 (0.98, 1)	1 (1, 1)	0.20 (0.17, 0.23)	0.36 (0.31, 0.41)
		3	0.99 (0.97, 1)	1 (0.99, 1)	0.21 (0.17, 0.23)	0.36 (0.31, 0.4)
(0.2, 0.4)	1	1	1 (0.98, 1)	1 (1, 1)	0.22 (0.18, 0.25)	0.40 (0.35, 0.47)
		2	1 (0.99, 1)	1 (1, 1)	0.22 (0.19, 0.25)	0.43 (0.37, 0.48)
		3	1 (0.98, 1)	1 (1, 1)	0.22 (0.18, 0.26)	0.41 (0.36, 0.47)
	2	1	1 (1, 1)	1 (1, 1)	0.22 (0.19, 0.26)	0.44 (0.39, 0.48)
		2	1 (1, 1)	1 (1, 1)	0.22 (0.19, 0.26)	0.44 (0.39, 0.48)
		3	1 (1, 1)	1 (1, 1)	0.22 (0.19, 0.26)	0.44 (0.39, 0.48)
(0.4, 0.2)	1	1	0.38 (0.18, 0.62)	0.60 (0.34, 0.81)	0.35 (0.31, 0.4)	0.26 (0.21, 0.31)
		2	0.39 (0.17, 0.64)	0.59 (0.33, 0.81)	0.35 (0.31, 0.4)	0.27 (0.21, 0.32)
		3	0.39 (0.18, 0.62)	0.60 (0.34, 0.81)	0.35 (0.31, 0.4)	0.27 (0.21, 0.31)
	2	1	0.51 (0.25, 0.77)	0.68 (0.41, 0.89)	0.34 (0.3, 0.39)	0.29 (0.23, 0.35)
		2	0.55 (0.27, 0.8)	0.71 (0.44, 0.9)	0.34 (0.29, 0.39)	0.3 (0.24, 0.35)
		3	0.52 (0.26, 0.76)	0.68 (0.41, 0.88)	0.34 (0.29, 0.39)	0.29 (0.23, 0.35)

[a]Each entry is the simulation average median, with first and third quantiles in parentheses. $\pi_{1,2}(\epsilon) = \Pr(\theta_1 - \epsilon < \theta_2 \mid data)$ for $\epsilon = .05$ or .10.

Table 2 gives an assessment of the sensitivity of posterior quantities to the assumed numerical prior correlation ρ in Method 1. In Table 2, the same posterior quantities considered in Table 1 are given for assumed $\rho = -.50, 0, +.50$, with physicians weighted equally, for the case where the true response probabilities are $(\bar{\theta}_1, \bar{\theta}_2) = (0.4, 0.2)$. Compared to $\rho = 0$, assuming either $\rho = -.50$ or $+.50$ increases the posterior equivalence probabilities $\pi^E_{12}(.05)$ and $\pi^E_{12}(.10)$ by .04 and changes the lower bound of the 95% posterior CI by .01. It thus appears that posterior inferences are relatively insensitive to ρ within this range.

To summarize the substantive conclusions of the simulations in Table 1, in the three cases where the true parameters are $(\bar{\theta}_1, \bar{\theta}_2) = (.50, .50)$, $(.20, .30)$, or $(.20, .40)$, all combinations of Method and physician weights give very large posterior probabilities that $\bar{\theta}_2$ is either .05- or .10-equivalent to $\bar{\theta}_1$. In these three cases, Method 2, which incorporates physician covariates and reflects the prior more strongly by giving larger posterior medians for $\bar{\theta}_2$ than for $\bar{\theta}_1$. When the true probabilities are (.40, .20), i.e. they are reversed so that treatment 1 is superior to treatment 2, the posterior equivalence probabilities are not large, $\bar{\theta}_1$ has larger posterior medians than $\bar{\theta}_2$, and most of the pairs of posterior 95% CIs either are disjoint or overlap slightly.

5 Sensitivity to prior bias and informativeness

The bins-and-chips elicitation method, and thus our proposed methodology, relies on first obtaining a sample of physician experts. In our application, choosing this sample was motivated by the desire to obtain priors from individuals who had experience treating idiopathic nephrotic syndrome with cyclophosphamide and MMF. This process was constrained by the fact that the pool of such physicians was limited. In general, practicing physicians typically have strong opinions, based on their experiences treating patients, and this was the case with our sample of experts. A key issue is that what is regarded as undesirable bias from one viewpoint may be regarded as valuable prior information from another. If such expert opinion is regarded as bias rather than useful prior information, then the negative connotation of the word "bias" reflects this viewpoint. This might motivate the desire for a

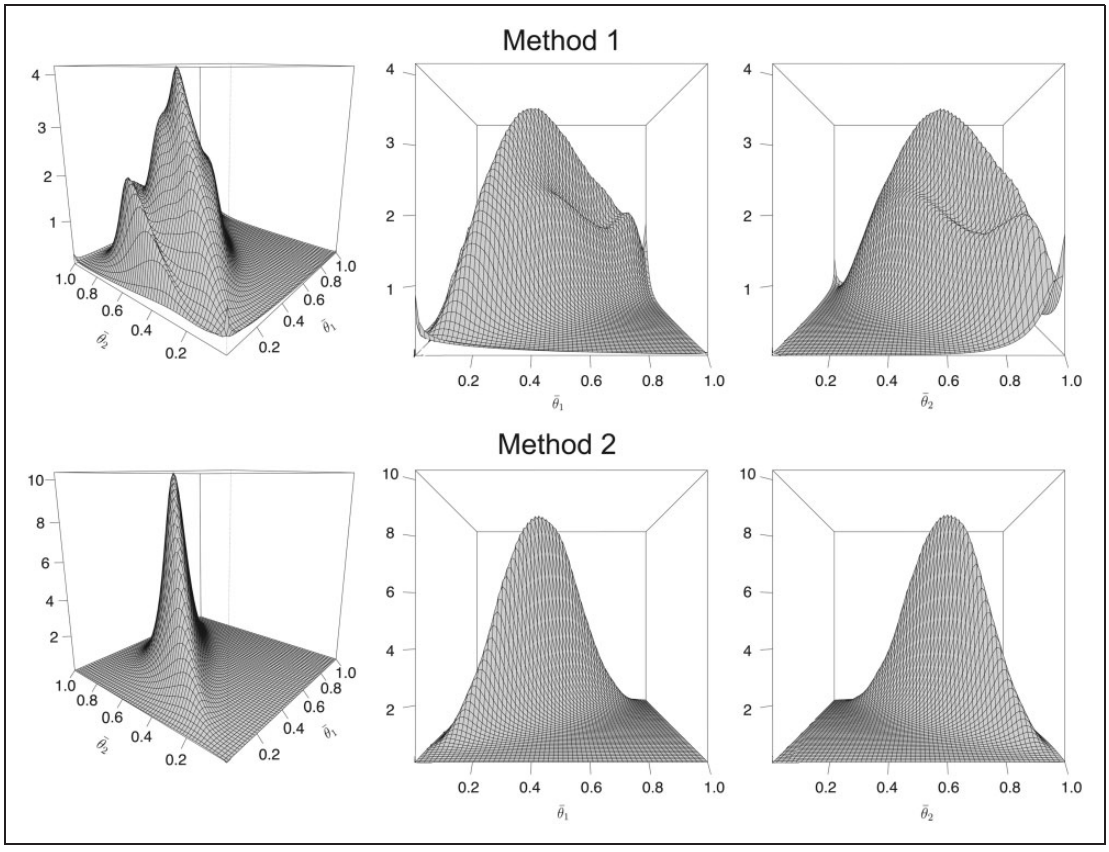

Figure 3. Three-dimensional plots of prior distributions of $(\bar{\theta}_1, \bar{\theta}_2)$ using Method 1 (top row) and Method 2 (bottom row), with equal physician weights. For Method 2, the covariates $\boldsymbol{X}_k$ were the logarithm of the number of year as pediatrician, the logarithm of the average number of patients consulted per year, and a binary indicator of whether the physician had training in clinical trial methodology.

Table 2. Sensitivity of Method 1 to the assumed numerical correlation ρ between the physician latent effects $(\epsilon_{k,\mu}, \epsilon_{k,\gamma})$.[a]

			Posterior median (25th, 75th percentiles)	
ρ	$\pi^E_{12}(0.05)$	$\pi^E_{12}(0.10)$	$\bar{\theta}_1$	$\bar{\theta}_2$
–.50	0.43 (0.21, 0.67)	0.64 (0.38, 0.84)	0.35 (0.30, 0.40)	0.27 (0.22, 0.31)
0	0.39 (0.18, 0.63)	0.60 (0.34, 0.81)	0.35 (0.31, 0.40)	0.27 (0.21, 0.31)
+.50	0.43 (0.21, 0.67)	0.64 (0.38, 0.84)	0.35 (0.30, 0.40)	0.27 (0.22, 0.31)

[a]Simulations are of a 70-subject trial with equally weighted physicians, for true $(\bar{\theta}_1, \bar{\theta}_2) = (0.4, 0.2)$, evaluating the same posterior quantities as in Table 1.

sample of independent or impartial experts. However, in practice, such a sample might be difficult or impossible to obtain for physicians treating a particular rare disease. Such a viewpoint thus might motivate the use of a more conventional vague or noninformative prior, with no elicitation at all, or a frequentist analysis of the data. If, instead, elicited expert opinion is regarded as valuable information, then the resulting prior is a valid basis for performing a Bayesian analysis. This is the motivation for our proposed methodology. Introducing the subjectivity of expert opinion via the prior in the analysis does not mean that our method is not objective, however. While the use of subjective probabilities to quantify prior uncertainty is a major criticism of Bayesian inference, from the Bayesian point of view, ''subjective'' does not mean arbitrary; just as from the frequentist point of view, ''objective'' does not mean without assumptions.

We address this issue as follows. In the Bayesian setting, it is always worthwhile to assess the influence of one's prior on posterior inferences. In the present setting, the following approach for constructing a set of alternative priors, and corresponding posterior inferences, provides a practical way to do this. These priors provide a set of intermediate approaches between the use of the informative prior that we have constructed and a noninformative prior.

To perform a sensitivity analysis of posterior inferences to the prior, for Method 1, we first define the expert-specific location parameters $\xi_{j,k} = \upsilon_{j,k,\mu}$ and precision parameters $\chi_{j,k} = \upsilon_{j,k,\gamma}$. For Method 2, we define these to be $\xi_{j,k} = \upsilon_{j,\mu} + \boldsymbol{\beta}_\mu \boldsymbol{X}_k$, and $\chi_{j,k} = \upsilon_{j,\gamma} + \boldsymbol{\beta}_\gamma \boldsymbol{X}_k$. Denoting $\boldsymbol{\xi}_j = (\xi_{j,1}, \cdots, \xi_{j,K})$ and $\boldsymbol{\chi}_j = (\chi_{j,1}, \cdots, \chi_{j,K})$ for $j = 1, 2$, we define the following two transformations of $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ to adjust prior location (bias) and of $\boldsymbol{\chi} = (\boldsymbol{\chi}_1, \boldsymbol{\chi}_2)$ to adjust prior informativeness. For location, one may specify several fixed values of $0 \leq \phi \leq 1$, which is a shift sensitivity parameter that replaces $\boldsymbol{\xi}_2$ with $(1 - \boldsymbol{\phi})\boldsymbol{\xi}_1 + \boldsymbol{\phi}\boldsymbol{\xi}_2$. Since the prior bias is determined by $\boldsymbol{\xi}_2 - \boldsymbol{\xi}_1$, specifying $\phi = 1$ gives the bias of the unadjusted prior, but as $\phi \downarrow 0$, the prior bias $\rightarrow 0$. The maximum shift is obtained at $\phi = 0$, where $\boldsymbol{\xi}_2 = \boldsymbol{\xi}_1$, so the prior bias is 0. For precision, we specify several fixed values of $0 < \lambda \leq 1$, a scale sensitivity parameter that replaces $\boldsymbol{\chi}_j$ with $\lambda\boldsymbol{\chi}_j$ for both $j = 1$ and $j = 2$. Thus, $\lambda = 1$ returns the original prior precision, while both priors become less informative as $\lambda \downarrow 0$. Thus, for each specified pair of fixed values of (ϕ, λ) used to transform the prior, computing the joint prior using either (5) or (6), one may compare particular posterior quantities obtained to those obtained for each (ϕ, λ) pair with the corresponding posterior values for $(\phi, \lambda) = (1,1)$, which gives the untransformed prior. In practice, to perform an analysis of sensitivity to both prior bias and prior informativeness when analyzing a given data set, one may choose a small number of values of (ϕ, λ), and assess each of several key posterior quantities. We will illustrate this below.

While any posterior quantities may be computed, the following are useful. In settings where the goal is to determine whether one treatment provides a given fixed improvement $\delta_\theta > 0$ over the other in the key parameter, one may compute the posterior probability of at least δ superiority of θ_2 over θ_1, $\pi^S_{1,2,\phi,\delta}(\delta) = P_{\phi,\lambda}(\theta_1 + \delta_\theta < \theta_2 \mid data)$, or possibly the symmetric probability $P_{\phi,\lambda}(|\theta_1 - \theta_2| > \delta_\theta \mid data)$. To quantify equivalence in a trial where treatment 1 is the standard and treatment 2 is the experimental, a relevant posterior probability is the ϵ-equivalence probability $\pi^E_{1,2,\phi,\delta}(\epsilon)$, now also indexed by the prior transformation parameters (ϕ, λ). Another useful quantity may be a 95% posterior CI for $\theta_2 - \theta_1$, which we denote by $CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$.

For illustration of how a prior-to-posterior sensitivity analysis may be done using a set of priors constructed from a matrix of (ϕ, λ) pairs, we first simulated one data set very similar to the data obtained from the NEPHROMYCY trial by setting fixed values $\theta_{1,true} = \theta_{2,true} = 0.4$, and simulating binary responses for 35 patients in each arm. In the simulated arm 1, the treatment was efficacious for 14 (40%) of 35 children and for 16 (45.7%) of 35 children in arm 2. For each (ϕ, λ), we constructed the prior using Method 1 with equal physician weights and computed three posterior values for this data set. Table 3 presents the posterior values for the 12 (λ, ϕ) pairs obtained from $\phi = 1$, 0.5, 0 and $\lambda = 1$, .75, .50, .25. In Table 3, the posterior probabilities of .05-equivalence $\pi^E_{1,2,\phi,\delta}(.05) = \Pr(\theta_1 - .05 < \theta_2 | data, \phi, \delta)$ and .15-superiority $\pi^S_{1,2,\phi,\delta}(.15) = \Pr(\theta_1 + .15 < \theta_2 | data, \phi, \delta)$, and the posterior 95% CI $CI_{95,\phi,\lambda}(\theta_2 - \theta_1 \mid data)$, are reported in each cell.

Table 3. Prior-to-posterior sensitivity analyses performed on a 70-patient data set with 14/35 responses in arm 1 and 16/35 responses in arm 2.[a]

		$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.2$
$\phi = 1$	$\pi^E_{1,2,\phi,\delta}(.05)$	0.92	0.88	0.84	0.77
	$\pi^S_{1,2,\phi,\lambda}(.15)$	0.24	0.22	0.21	0.18
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.09, 0.27)	(–0.13, 0.26)	(–0.17, 0.27)	(–0.20, 0.27)
$\phi = 0.5$	$\pi^E_{1,2,\phi,\lambda}(.05)$	0.77	0.76	0.75	0.75
	$\pi^S_{1,2,\phi,\delta}(.15)$	0.17	0.17	0.15	0.16
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.18, 0.26)	(–0.20, 0.26)	(–0.19, 0.26)	(–0.19, 0.25)
$\phi = 0$	$\pi^E_{1,2,\phi,\lambda}(.05)$	0.77	0.75	0.76	0.74
	$\pi^S_{1,2,\phi,\lambda}(.15)$	0.16	0.15	0.15	0.15
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.19, 0.26)	(–0.21, 0.26)	(–0.20, 0.26)	(–0.20, 0.25)

[a]The prior was constructed using Method 1 and equal physician weights and was transformed for each (ϕ, λ) pair, and the posterior quantities $\pi^E_{1,2,\phi,\lambda}(.05) = \Pr(\theta_1 - .05 < \theta_2 \mid data, \phi, \lambda)$, $\pi^S_{1,2,\phi,\lambda}(.15) = \Pr(\theta_1 + .15 < \theta_2 \mid data, \phi, \lambda)$ and $CI_{95,\phi,\lambda}(\theta_2 - \theta_1 \mid data)$ then were computed.

Table 4. Prior-to-posterior sensitivity analyses repeated from Table 3, but using a new mixture prior computed from the 17 actual experts plus 17 synthetic experts obtained as a bootstrap sample from the set of $(\mu_{1,k}, \gamma_{1,k}, \mu_{2,k}, \gamma_{2,k})$ values.[a]

		$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.2$
$\phi = 1$	$\pi^E_{1,2,\phi,\delta}(.05)$	0.93	0.88	0.83	0.79
	$\pi^S_{1,2,\phi,\lambda}(.15)$	0.27	0.23	0.21	0.18
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.1, 0.28)	(–0.13, 0.27)	(–0.16, 0.27)	(–0.18, 0.27)
$\phi = 0.5$	$\pi^E_{1,2,\phi,\lambda}(.05)$	0.79	0.77	0.76	0.75
	$\pi^S_{1,2,\phi,\delta}(.15)$	0.16	0.16	0.15	0.15
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.18, 0.25)	(–0.19, 0.27)	(–0.2,0.25)	(–0.19, 0.26)
$\phi = 0$	$\pi^E_{1,2,\phi,\lambda}(.05)$	0.76	0.76	0.77	0.75
	$\pi^S_{1,2,\phi,\lambda}(.15)$	0.14	0.16	0.16	0.15
	$CI_{95,\phi,\lambda}(\theta_2 - \theta_1)$	(–0.20, 0.25)	(–0.20, 0.26)	(–0.19, 0.26)	(–0.19, 0.25)

[a]The 17 bootstrap sample values were jittered by adding independent $N(0, .5^2)$ noise to each logit$(\mu_{j,k})$ and each log$(\gamma_{j,k})$ before computing the new prior.

Table 3 shows that the posterior probability that the two treatments are 0.05-equivalent is 0.92 for the untransformed prior with $(\phi, \lambda) = (1,1)$. For no shift in prior location ($\phi = 1$), $\pi^E_{1,2,1,\delta}(.05)$ drops to 0.77 when the prior precision is reduced to 25% of its original value, i.e. $\lambda = .25$, which corresponds to a prior effective sample size of about $.25 \times 70 = 17.5$. For prior shift parameter $\phi = 0.5$ or 0, the value of $\pi^E_{1,2,\phi,\delta}(.05)$ drops to values in the narrow range 0.74 to 0.77 for all λ. The posterior probability $\pi^S_{1,2,\phi,\delta}(.15)$ of 0.15-superiority is 0.24 for the untransformed prior, and for $\phi = 1$, this drops most, to 0.18, as the precision λ is reduced from 1 to 0.25. In contrast, for the shifted priors obtained by $\phi = 0.5$ or 0, $\pi^S_{1,2,\phi,\delta}(.15)$ is insensitive to λ, taking on values 0.15 to 0.17 in these six cases. The upper limit on the posterior 95% CI for θ_2–θ_1 is insensitive to all (ϕ, λ) values, but for $\phi = 1$, the lower limit decreases from –0.09 to –0.20 as λ drops from 1 to 0.25, and otherwise is insensitive to (ϕ, λ) for $\phi = 0.5$ or 0.

A natural question is whether a larger sample of experts might provide a more reliable mixture prior, and thus alter posterior inferences. To address this, we drew a bootstrap sample of size 17 from the set of $(\mu_{1,k}, \gamma_{1,k}, \mu_{2,k}, \gamma_{2,k})$ values obtained from the beta distributions fit to the elicited histograms. We then jittered each of these new values by adding independent $N(0,.5^2)$ noise to each logit$(\mu_{j,k})$ and each log$(\gamma_{j,k})$. We treated the resulting values as transformed prior beta parameters from an additional sample of 17 synthetic experts. We combined these new parameters with the original parameters from the actual sample of experts and computed a new mixture prior, again using Method 1 with equal weights. The results are summarized in Table (4) All posterior quantities, including 95% CIs, are very similar to the corresponding values in Table (3). It thus appears, in this example, that doubling the number of experts does not alter one's posterior inferences substantively in terms of either location or variability, at least if the new experts are similar to the original experts.

6 Discussion

We have provided a methodology for constructing informative parametric mixture priors, based on histograms of key treatment parameters elicited from expert physicians by applying the graphical method of Johnson et al.[6] Our motivation was the desire to deal with settings where the sample size of a randomized trial is not large enough to obtain confirmatory results using conventional statistical methods, but physicians experienced with the disease and treatments are available to provide their opinions. Because we give methods that either do or not incorporate physician covariates in the marginal physician-specific priors, and also three different ways to weight physicians when computing the overall mixture prior, there are a total of six different versions of the methodology. Since posterior quantities appear to be insensitive to how physicians are weighted when computing the mixture, however, in practice, it seems best simply to weight the physicians equally. While we have focused on the case of a binary outcome, the approach may be adapted to real-valued or time-to-event outcome data in a straightforward manner. Because incorporating expert opinion into the prior used for a Bayesian analysis is inherently controversial, we also have provided an explicit method for constructing a set of alternative priors, each obtained by applying a location shift and a change in precision. This provides a framework for performing a sensitivity analysis in an explicit way to use as a basis for making informed conclusions about the comparative benefits of the two treatments.

The methodology proposed could be extended to accommodate multiarm trials, with $K > 2$ parameters. This would require one to elicit K histograms from each physician, however. Since it seems likely that some experts would be familiar with only a subset of the K treatments, and thus different experts might provide priors on different subvectors of $(\bar{\theta}_1, \cdots, \bar{\theta}_K)$, the problem of weighting the experts when constructing a mixture prior might not be entirely straightforward.

Computer script, written in R, which was used to implement the methodology, is available as a Supplementary file.

Acknowledgements

The authors thank Georges Deschłnes for assistance in preparing the elicitation questionnaire.

Authors' note

A list of the physician experts who participated in the elicitation is given in the supplementary materials.

Authors' contributions

Peter Thall and Moreno Ursino made equal contributions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: Peter Thall's research was supported by NCI grant 5-R01-CA083932 and grant IDEX from the Universite Sorbonne Paris Cite (2013, project 24). Moreno Ursino and Sarah Zohar were funded by the InSPiRe Project of the European Union Seventh Framework Programme for Research, Technological Development, and Demonstration under grant agreement FP HEALTH 2013-602144.

Supplemental material

Supplemental material for this article is available online.

References

1. Fakhouri F, Bocquet N, Taupin P, et al. Steroid-sensitive nephrotic syndrome: from childhood to adulthood. *Am J Kidney Dis* 2003; **41**: 550–557.
2. Rüth EM, Kemper MJ, Leumann EP, et al. Children with steroid-sensitive nephrotic syndrome come of age: long-term outcome. *J Pediatr* 2005; **147**: 202–207.
3. Pennisi AJ, Grushkin CM and Lieberman E. Cyclophosphamide in the treatment of idiopathic nephrotic syndrome. *Pediatr* 1976; **57**: 948–951.
4. Bagga A, Hari P, Moudgil A, et al. Mycophenolate mofetil and prednisolone therapy in children with steroid-dependent nephrotic syndrome. *Am J Kidney Dis* 2003; **42**: 1114–1120.
5. Barletta GM, Smoyer WE, Bunchman TE, et al. Use of mycophenolate mofetil in steroid-dependent and-resistant nephrotic syndrome. *Pediatr Nephrol* 2003; **18**: 833–837.
6. Johnson SR, Tomlinson GA, Hawker GA, et al. A valid and reliable belief elicitation method for bayesian priors. *J Clin Epidemiol* 2010; **63**: 32: 370–383.
7. Savage LJ. Elicitation of personal probabilities and expectations. *J Am Stat Assoc* 1971; **66**: 783–801.
8. Chaloner KM and Duncan GT. Assessment of a beta prior distribution: Pm elicitation. *Statistician* 1983; **32**: 174–180.
9. Kadane J and Wolfson LJ. Experiences in elicitation. *J R Stat Soc: Series D (Statistician)* 1998; **47**: 3–19.
10. O'Hagan A. Eliciting expert beliefs in substantial practical applications. *J R Stat Soc: Series D (Statistician)* 1998; **47**: 21–35.
11. Chaloner K and Rhame FS. Quantifying and documenting prior beliefs in clinical trials. *Stat Med* 2001; **20**: 581–600.

12. Kuhnert PM, Martin TG and Griffiths SP. A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecol Lett* 2010; **13**: 900–914.
13. Spiegelhalter DJ, Harris NL, Bull K, et al. Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. *J Am Stat Assoc* 1994; **89**: 435–443.
14. Tan SB, Chung YFA, Tai BC, et al. Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. *Contr Clin Trials* 2003; **24**: 110–121.
15. Hiance A, Chevret S and Lévy V. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. *J Clin Epidemiol* 2009; **62**: 431–437.
16. DuMouchel W. A Bayesian model and graphical elicitation procedure for multiple comparisons. *Bayesian Stat* 1988; **3**: 127–145.
17. Chaloner K, Church T, Louis TA, et al. Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* 1993; **42**: 341–353.
18. Clemen RT and Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Anal* 1999; **19**: 187–203.
19. Moatti M, Zohar S, Facon T, et al. Modeling of experts divergent prior beliefs for a sequential phase III clinical trial. *Clin Trials* 2013; **10**: 505–514.
20. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain judgements: eliciting experts' probabilities*. Chichester: John Wiley & Sons, 2006.
21. Albert I, Donnet S, Guihenneuc-Jouyaux C, et al. Combining expert opinions in prior elicitation. *Bayesian Anal* 2012; **7**: 503–532.
22. Wahed AS and Thall PF. Evaluating joint effects of induction–salvage treatment regimes on overall survival in acute leukaemia. *J Royal Stat Soc: Series C (Appl Stat)* 2013; **62**: 67–83.