Bayesian Analysis (2012) **7**, Number 3, pp. 591–614

Prior Effective Sample Size in Conditionally Independent Hierarchical Models

Satoshi Morita[*], Peter F. Thall[†] and Peter Müller[‡]

Abstract. Prior effective sample size (ESS) of a Bayesian parametric model was defined by Morita, et al. (2008, *Biometrics*, **64**, 595-602). Starting with an ε-information prior defined to have the same means and correlations as the prior but to be vague in a suitable sense, the ESS is the required sample size to obtain a hypothetical posterior very close to the prior. In this paper, we present two alternative definitions for the prior ESS that are suitable for a conditionally independent hierarchical model. The two definitions focus on either the first level prior or second level prior. The proposed methods are applied to important examples to verify that each of the two types of prior ESS matches the intuitively obvious answer where it exists. We illustrate the method with applications to several motivating examples, including a single-arm clinical trial to evaluate treatment response probabilities across different disease subtypes, a dose-finding trial based on toxicity in this setting, and a multicenter randomized trial of treatments for affective disorders.

Keywords: Bayesian hierarchical model, Conditionally independent hierarchical model, Computationally intensive methods, Effective sample size, Epsilon-information prior

1 Introduction

Recently, a definition for the effective sample size (ESS) of a given prior $\pi(\boldsymbol{\theta})$ with respect to a sampling model $p(Y \mid \boldsymbol{\theta})$ was proposed by Morita, Thall, and Müller (2008) (MTM). The ESS provides an easily interpretable index of the informativeness of a prior with respect to a given likelihood. The approach is to first define an ε-information prior $\pi_0(\boldsymbol{\theta})$ having the same means and correlations as $\pi(\boldsymbol{\theta})$ but being vague in a suitable sense, and then define the ESS to be the sample size n of hypothetical outcomes $\mathbf{Y}_n = (Y_1, \cdots, Y_n)$ that, starting with $\pi_0(\boldsymbol{\theta})$, yields a hypothetical posterior $\pi_n(\boldsymbol{\theta} \mid \mathbf{Y}_n)$ very close to $\pi(\boldsymbol{\theta})$. MTM define the distance between $\pi(\boldsymbol{\theta})$ and $\pi_n(\boldsymbol{\theta} \mid \mathbf{Y}_n)$ in terms of the trace of the negative second derivative matrices of $\log\{\pi(\boldsymbol{\theta})\}$ and $\log\{\pi_n(\boldsymbol{\theta} \mid \mathbf{Y}_n)\}$. The ESS is defined as the interpolated value of n that minimizes this "prior-to-posterior" distance. While this definition is suitable for a wide range of models and applications, it fails for hierarchical models.

[*] Department of Biostatistics and Epidemiology, Yokohama City University Medical Center, Yokohama, Japan, smorita@yokohama-cu.ac.jp

[†] Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, rex@mdanderson.org

[†] Department of Mathematics, The University of Texas at Austin, Austin, TX, pmueller@math.utexas.edu

DOI:10.1214/12-BA720

In this paper, we propose two extensions of the definition for prior ESS that are applicable to two-stage conditionally independent hierarchical models (CIHMs) (Kass and Steffey, 1989). Our approach is pragmatic in that the ESS can always be evaluated either analytically or by using a simulation-based approach. We validate the definitions by verifying that they match the intuitively obvious answers in important special cases where such answers exist. We focus on the class of CIHMs due to their practical importance in settings where data are collected from exchangeable subgroups, such as study centers, schools, cities, etc. Important areas of application include meta-analysis (cf. Berry and Stangl, 1996; Berlin and Colditz, 1999), and clinical trial design (cf. Thall *et al.*, 2003).

Moreover, we restrict attention to CIHMs as the most commonly used versions of hierarchical models. For more complex hierarchical models, one might report prior ESS values for appropriate sub-models, although it might become less meaningful to report an overall ESS.

A two-level CIHM for K subgroups is defined as follows. Let $\mathbf{Y}_k = (Y_{k,1}, \ldots, Y_{k,n_k})$ denote the vector of outcomes for sub-group k and let $\mathcal{Y}_M = (\mathbf{Y}_1, ..., \mathbf{Y}_K)$, with the K vectors assumed to be distributed independently conditional on hyperparameters. We use $f(\cdot)$ generically to indicate the sampling model of observable data, which may be a single variable $Y_{k,j}$, a vector $\mathbf{Y}_k$, or the vector $\mathcal{Y}_M$ of all $n_1 + \cdots + n_K$ observations. The nature of the argument will clarify the specific meaning of $f(\cdot)$. In the first level, $\mathbf{Y}_k$ follows distribution $f(\mathbf{Y}_k \mid \boldsymbol{\theta}_k)$. In the second level, the subgroup-specific parameters $\boldsymbol{\theta}$=$(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ are assumed to be i.i.d. with prior $\pi_1(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}})$, where the hyperparameter $\tilde{\boldsymbol{\theta}}$ has a hyperprior $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ with known $\boldsymbol{\phi}$. The model is summarized in equation (1).

$$\begin{array}{lrcl} \text{Sampling model} & f(\mathcal{Y}_M \mid \boldsymbol{\theta}) & = & \prod_{k=1}^K f(\mathbf{Y}_k \mid \boldsymbol{\theta}_k) \\ \text{(Level 1) Prior} & \pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) & = & \prod_{k=1}^K \pi_1(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}}) \\ \text{(Level 2) Hyperprior} & \pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}). & & \end{array} \quad (1)$$

A common example of a CIHM (1) is a conjugate normal/inverse χ^2-normal-normal model. Let Inv-$\chi^2(\nu, S)$ denote a scaled inverse χ^2 distribution with ν degrees of freedom, mean $\frac{\nu S}{\nu-2}$ for $\nu > 2$, and variance $\frac{2\nu^2 S^2}{(\nu-2)^2(\nu-4)}$ for $\nu > 4$. This model has a normal sampling distribution $Y_{k,i} \mid \theta_k \sim \mathrm{N}(\theta_k, \sigma^2)$ with known σ^2, where $\mathrm{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Independent conjugate normal priors $\theta_k \mid \tilde{\mu}, \tilde{\gamma}^2 \sim \mathrm{N}(\tilde{\mu}, \tilde{\gamma}^2)$ on the location parameters $\theta_1, \cdots, \theta_K$ are assumed, with a normal/inverse χ^2 hyperprior $\tilde{\mu} \mid \mu_\phi, \tau_\phi^2 \sim \mathrm{N}(\mu_\phi, \tau_\phi^2)$ and $\tilde{\gamma}\ ^2 \mid \nu_\phi, S_\phi \sim$ Inv-$\chi^2(\nu_\phi, S_\phi)$. Once the methodology is established, we will explain how to compute ESS when σ^2 is not assumed to be known but rather is random with its own prior.

To compute an ESS under a CIHM, we will consider the following two cases, which address different inferential objectives. In case 1, the target is the marginalized prior,

$$\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}) = \int \pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) \pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}) d\tilde{\boldsymbol{\theta}}. \quad (2)$$

An example is a setting where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are treatment effects in K different disease subtypes and they are the parameters of primary interest. In case 2, the target prior is

$$\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}). \tag{3}$$

For example, this would arise if an overall treatment effect $\tilde{\boldsymbol{\theta}}$ obtained by averaging over K clinical centers in a multi-center clinical trial is the parameter of primary interest. We propose two definitions for the ESS under a CIHM, one for each case, allowing the possibility that both types of ESS may be of interest in a given analysis. For later reference we define the marginal likelihood

$$f_1(\mathbf{Y}_k \mid \tilde{\boldsymbol{\theta}}) = \int f(\mathbf{Y}_k \mid \boldsymbol{\theta}_k)\pi_1(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}})d\boldsymbol{\theta}_k \tag{4}$$

by integrating with respect to the level 1 prior.

Section 2 presents motivating examples. We briefly summarize the MTM formulation in Section 3. The two definitions of ESS in CIHMs and accompanying computational methods are given in Section 3. In Section 4 we compute the ESS for the three motivating examples. In Section 5 we discuss some standard CIHMs, and we close with a brief discussion in Section 6.

2 Motivating examples

2.1 A Single-Arm Sarcoma Trial

Thall et al. (2003) present a design for a single-arm phase II trial to examine the efficacy of the targeted drug imatinib for sarcoma, a disease with many subtypes. Since sarcomas are uncommon, the goal was to construct a design that allowed the efficacy of imatinib to be evaluated in $K = 10$ sarcoma subtypes. This was achieved by assuming the following CIHM, where the treatment effects differ across subtypes. The parameters of primary interest were the subtype-specific tumor response probabilities, $\xi_1, \cdots, \xi_{10}$. Let $\text{Ga}(a_\phi, b_\phi)$ denote a gamma distribution with mean a_ϕ/b_ϕ and variance a_ϕ/b_ϕ^2. Denoting $\theta_k = \log\{\xi_k/(1-\xi_k)\}$, it was assumed that $\theta_1, \ldots, \theta_{10}$ were i.i.d. $\text{N}(\tilde{\mu}, \tilde{\gamma}^{-1})$ and that $\tilde{\mu}$ and the precision parameter $\tilde{\gamma}$ followed independent normal and gamma hyperpriors, respectively. Elicitation of prior probabilities characterizing association between pairs of ξ_k's yielded the hyperpriors $\tilde{\mu} \sim \text{N}(-1.386, 10)$ and $\tilde{\gamma} \sim \text{Ga}(2, 20)$, so that $E(\tilde{\gamma}) = .10$ and $\text{var}(\tilde{\gamma}) = .005$. In summary, the trial design assumed the following model:

$$\begin{array}{lrcl}
\text{Sampling model} & \mathbf{Y}_{k,m} \mid \theta_k & \sim & \text{Bin}(m, \xi_k) \text{ indep. for all } k \\
\text{Prior} & \theta_k \mid \tilde{\mu}, \tilde{\gamma} & \sim & \text{N}(\tilde{\mu}, \tilde{\gamma}^{-1}) \text{ i.i.d. for all } k \\
\text{Hyperpriors} & \tilde{\mu} & \sim & \text{N}(-1.386, 10) \\
 & \tilde{\gamma} & \sim & \text{Ga}(2, 20).
\end{array} \tag{5}$$

Thall et al. (2003) used the marginal posterior probability $\Pr(\xi_k > 0.30|\mathcal{Y})$ to define an early stopping criterion in disease subtype k, which was computed based on the posterior $\pi(\boldsymbol{\theta} \mid \mathcal{Y})$ under (5). Thus, 10 stopping rules were applied, one for each subtype. Note that $\mathcal{Y}$ included the data from all ten subtypes in order to exploit the association among the θ_k's induced by the hierarchical model. This rule was first applied after observing a minimum of eight patients in each disease subtype, and subsequently at sample sizes of 17, 23, and 30 patients. Thus, an overly informative prior, for example, with a prior ESS ≥ 40 might be considered inappropriate since the prior, rather than patient response data, would dominate early termination decisions. Because both a prior and hyperprior were specified, the methods in MTM are not applicable, and it is not obvious how to determine the ESS of this model. We will show below that the ESS can be determined for this model using an approach that is coherent in the sense that it gives the intuitively obvious answer in cases where the ESS exists.

2.2 A CRM Dose-finding Trial for Multiple Patient Subgroups

As an extension of the hierarchical model (HM) in (5) we consider a phase I dose-finding trial with multiple patient subgroups using a model-based design. We assume an implementation that generalizes the continual reassessment method (CRM) (O'Quigley *et al.*, 1990). Suppose that there are $K = 4$ subgroups ($k = 1, \ldots, 4$) with population proportions $(.40, .30, .20, .10)$. Each patient in each subgroup receives one of six doses, 100, 200, 300, 400, 500, 600, denoted by $d_1, \ldots, d_6$, with standardized doses $x_z = \log(d_z) - 1/6\sum_{l=1}^{6}\log(d_l)$. The outcome variable is the indicator $Y_{k,i} = 1$ if the i^{th} patient in subgroup k suffers toxicity, 0 if not. The probability of toxicity in subgroup k under dose x_i is denoted by $p_k(x_i, \alpha_k, \beta_k) = \Pr(Y_{k,i} = 1 \mid x_i, \alpha_k, \beta_k)$ with logit $\{p_k(x_i, \alpha_k, \beta_k)\} = \alpha_k + \beta_k x_i$, for $k = 1, 2, 3, 4$. We have a CRM-type goal of finding the "optimal" dose x_k^* in each subgroup k. Optimal is defined as the posterior mean of $p_k(x_k^*)$ being closest to some fixed target p^*. The maximum sample size is 36, with the cohort size of 1, starting at the lowest dose d_1, and not skipping a dose level when escalating, with target toxicity probability $p^* = .30$. The parameters of primary interest are $\boldsymbol{\theta}_k = (\alpha_k, \beta_k)$, $k = 1, 2, 3, 4$. It is assumed that $\alpha_1, \ldots, \alpha_4$ and $\beta_1, \ldots, \beta_4$ are i.i.d. $\mathrm{N}(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $\mathrm{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$, respectively, and that $\tilde{\mu}_\alpha$ and $\tilde{\mu}_\beta$ follow independent normal hyperpriors. For the variance hyperparameters $\tilde{\sigma}_\alpha^2$ and $\tilde{\sigma}_\beta^2$, following Gelman (2006, Section 4.3) we assume that $\tilde{\sigma}_\alpha$ and $\tilde{\sigma}_\beta$ are uniform on $[0, U_\phi]$. Denoting the dose assigned to the i^{th} patient by $x_{[i]}$, in summary we assume

$$\begin{array}{lrcl}
\text{Sampling model} & Y_{k,i} \mid \boldsymbol{\theta}_k, x_{[i]} & \sim & \text{Bernoulli}(p_k(x_{[i]}, \boldsymbol{\theta}_k)) \text{ indep. for all } k \\
\text{Prior} & \alpha_k \mid \tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2 & \sim & \mathrm{N}(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2) \text{ i.i.d. for all } k \\
 & \beta_k \mid \tilde{\mu}_\beta, \tilde{\sigma}_\beta^2 & \sim & \mathrm{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2) \text{ i.i.d. for all } k \\
\text{Hyperpriors} & \tilde{\mu}_\alpha \mid \mu_{\alpha,\phi}, \sigma_{\alpha,\phi}^2 & \sim & \mathrm{N}(\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}^2) \\
 & \tilde{\mu}_\beta \mid \mu_{\beta,\phi}, \sigma_{\beta,\phi}^2 & \sim & \mathrm{N}(\mu_{\beta,\phi}, \sigma_{\beta,\phi}^2) \\
 & \tilde{\sigma}_\alpha, \tilde{\sigma}_\beta \mid U_\phi & \sim & \mathrm{U}(0, U_\phi).
\end{array} \qquad (6)$$

We will later, in Section 3.4, discuss how the ESS in this example depends not only on the assumed probability model and hyperparameters, but also on design choices like the adaptive dose-finding algorithm, the population proportions, etc. Note that we assume that $(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$ are independent, in order to have a reasonably parsimonious model. We can use elicited information to solve for the hyperprior means $\mu_{\alpha,\phi}$ and $\mu_{\beta,\phi}$, as follows. Given the standardized doses, those hyperprior means are calculated based on the elicited values $\mathrm{E}\{p_k(x_2 = -.403)\} = .25$ at the second lowest dose and $\mathrm{E}\{p_k(x_5 = .513)\} = .75$ at the second highest dose for all the subgroups. These give $\mu_{\alpha,\phi}$ = -0.131 and $\mu_{\beta,\phi} = 2.398$. We will evaluate the ESS under several combinations of $(\sigma^2_{\alpha,\phi}, \sigma^2_{\beta,\phi}, U_\phi)$ in a sensitivity analysis.

2.3 A Multicenter Randomized Trial

When analyzing data from a multicenter trial, it often is important to examine the inter-center variability of treatment effects, i.e., treatment-by-center interaction (Gray, 1994), since substantial variation among treatment effects across centers may cause a regulatory agency to question the generalizability of results obtained from such a trial before giving approval for a new therapy. As a third example, we consider a multicenter randomized clinical trial reported by Stangl (1996). The trial was carried out to examine the inter-center variability of the effect of imipramine hydrochloride for preventing the recurrence of depression. The primary outcome was time to the first recurrence of a depressive episode, denoted by $T_{jk,1}, ..., T_{jk,n_{jk}}$ for n_{jk} patients receiving treatment j at the k^{th} center, for $j = 1, 2$ and $k = 1, \ldots, K$. A total of 150 patients were enrolled in $K = 5$ centers. For each (j, k), the recurrence times $T_{jk,1}, \ldots, T_{jk,n_{jk}}$ were assumed to be i.i.d. exponentially distributed with recurrence rate θ_{jk}. Working with the transformed parameters $\zeta_k = \log(\theta_{1k}/\theta_{2k})$ and $\eta_k = \log(\theta_{2k})$, the priors were assumed to be $\zeta_k \sim$ i.i.d. $\mathrm{N}(\tilde{\mu}_\zeta, \tilde{\sigma}^2_\zeta)$ and $\eta_k \sim$ i.i.d. $\mathrm{N}(\tilde{\mu}_\eta, \tilde{\sigma}^2_\eta)$. The hyperparameter $\tilde{\sigma}^2_\zeta$ of the inter-center heterogeneity of the treatment effect log ratios is of primary interest in this example, while $\tilde{\sigma}^2_\eta$ represents the inter-center heterogeneity in the effect of the control treatment arm. Lognormal hyperpriors were assumed with $\tilde{\sigma}_\zeta \sim \mathrm{LN}(m_\phi, s^2_\phi)$, and $\tilde{\sigma}_\eta \sim \mathrm{LN}(-0.22, 1)$, where $\mathrm{LN}(\mu, \sigma^2)$ denotes the lognormal distribution of e^X for $X \sim N(\mu, \sigma^2)$. The model is summarized as follows:

$$\begin{array}{llcl}
\text{Sampling model} & T_{jk,i} \mid \theta_{jk} & \sim & \mathrm{Exp}(\theta_{jk}) \text{ indep. for } j = 1, 2, \text{ and all } k \\
\text{Priors} & \zeta_k \mid \tilde{\mu}_\zeta, \tilde{\sigma}^2_\zeta & \sim & \mathrm{N}(\tilde{\mu}_\zeta, \tilde{\sigma}^2_\zeta) \text{ i.i.d. for all } k \\
 & \eta_k \mid \tilde{\mu}_\eta, \tilde{\sigma}^2_\eta & \sim & \mathrm{N}(\tilde{\mu}_\eta, \tilde{\sigma}^2_\eta) \text{ i.i.d. for all } k \\
\text{Hyperpriors} & \tilde{\sigma}_\zeta \mid m_\phi, s^2_\phi & \sim & \mathrm{LN}(m_\phi, s^2_\phi), \\
 & \tilde{\sigma}_\eta & \sim & \mathrm{LN}(-0.22, 1^2), \\
 & \tilde{\mu}_\zeta, \tilde{\mu}_\eta & \sim & \mathrm{U}(-20, 20).
\end{array} \tag{7}$$

Stangl assumed two alternative sets of hyperparameters (m_ϕ, s^2_ϕ), to represent two types of prior belief on $\tilde{\sigma}^2_\zeta$ in a Bayesian sensitivity analysis. The first choice was $(m_\phi, s^2_\phi) = (-1.61, 0.50^2)$, which places substantial prior belief on smaller $\tilde{\sigma}^2_\zeta$, and the second was $(m_\phi, s^2_\phi) = (0, 0.50^2)$, which places prior weight on larger $\tilde{\sigma}^2_\zeta$. We will evaluate the ESS of each prior on $\tilde{\sigma}^2_\zeta$ under case 2 of our proposed methods.

Event time data, like the recurrence time, often includes extensive censoring. In the presence of censoring, the amount of information, and thus the ESS, depends on the number of observed events in addition to the sample size. The ESS computation in this example, therefore, needs to account for censoring cases which can occur depending on study duration. We will discuss details of the ESS computation in the presence of censoring and other relevant design details in Section 3.4.

In any CIHM, the prior choice is subject to two competing desiderata. On one hand, an informative hyperprior that expresses a belief of strong association among the θ_k's is needed to borrow strength across subpopulations. In some settings, it is appropriate to use an informative prior that reflects accurate and comprehensive prior knowledge. If the hyperprior is elicited from an area expert, then the ESS provides an easily understood numerical value that the expert may use, if desired, to modify his/her original elicited values. On the other hand, in some settings it is necessary to avoid excessively informative priors that may compromise the objectivity of one's conclusions. In practice, many arbitrary choices are made for technical convenience while formulating a model. A skeptical reviewer may like to quantify the prior information as being equivalent to a certain number of hypothetical observations, i.e., a prior ESS. Such a summary immediately allows a reader to judge the relative contributions of the prior and the data to the final conclusion.

3 Effective sample size in CIHMs

3.1 Prior Effective Sample Size in Non-hierarchical Models

In this subsection, we review and formalize the heuristic definition of ESS for non-hierarchical models given by MTM. We give formal definitions of the ε-information prior and a prior-to-posterior distance. The intuitive motivation for MTM's method is to mimic the rationale for why the ESS of a beta distribution, $Be(a, b)$, equals $a + b$. A binomial variable Y with binomial sample size n and success probability θ following a $Be(a, b)$ prior implies a $Be(a + Y, b + n - Y)$ posterior. Thus, saying that a given $Be(a, b)$ prior has ESS $m = a + b$ requires the implicit reasoning that the $Be(a, b)$ may be identified with a $Be(c + Y, d + m - Y)$ posterior arising from a previous $Be(c, d)$ prior having a very small amount of information. A simple way to formalize this is to set $c + d = \varepsilon$ for an arbitrarily small value $\varepsilon > 0$ and solve for $m = a + b - (c + d) = a + b - \varepsilon$.

In a generic, non-hierarchical model, let $f(Y \mid \boldsymbol{\theta})$ denote the distribution of a random, possibly vector-valued outcome Y and $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ the prior on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, with hyperparameters $\tilde{\boldsymbol{\theta}}$. The definition of the ESS of $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ given $f(Y \mid \boldsymbol{\theta})$ requires the notions of an ε-information prior and the distance between the prior and a hypothetical posterior corresponding to a sample of a given size used to update an ε-information prior. The following definition formalizes the heuristic definition given by MTM. Denote $\sigma^2_{\pi,j} = \text{var}_\pi(\theta_j)$.

Definition of epsilon-information prior: Let $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ be a prior with $\sigma^2_{\pi,j} < \bar{V}_j(\tilde{\boldsymbol{\theta}}) \leq \infty$, where $\bar{V}_j(\tilde{\boldsymbol{\theta}})$ is a fixed bound, for all $j = 1, ..., d$. Given arbitrarily small $\varepsilon > 0$, the prior $\pi_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ defined on the same domain and in the same parametric family as $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ is an *ε-information prior* if $\mathrm{E}_{\pi_0}(\boldsymbol{\theta}) = \mathrm{E}_{\pi}(\boldsymbol{\theta})$, $\mathrm{Corr}_{\pi_0}(\theta_j, \theta_{j'}) = \mathrm{Corr}_{\pi}(\theta_j, \theta_{j'})$ for all $j \neq j'$ and

$$\left\{ \frac{\bar{V}_j(\tilde{\boldsymbol{\theta}}) - \sigma^2_{\pi_0,j}}{\bar{V}_j(\tilde{\boldsymbol{\theta}}) - \sigma^2_{\pi,j}} \right\} \frac{\sigma^2_{\pi,j}}{\sigma^2_{\pi_0,j}} < \varepsilon \qquad \text{for } j = 1, ..., d. \tag{8}$$

In the case $\bar{V}_j(\tilde{\boldsymbol{\theta}}) = \infty$, condition (8) reduces to $\sigma^2_{\pi,j}/\sigma^2_{\pi_0,j} < \varepsilon$. Thus, in this case one may simply choose $\tilde{\boldsymbol{\theta}}_0$, subject to the constraints on the means and correlations, so that $\sigma^2_{\pi_0,j}$ is large enough to ensure that this inequality holds. When $\bar{V}_j(\tilde{\boldsymbol{\theta}}) < \infty$, note that (8) can be written in the form $\{\bar{V}_j(\tilde{\boldsymbol{\theta}})/\sigma^2_{\pi_0,j} - 1\} < \varepsilon\{\bar{V}_j(\tilde{\boldsymbol{\theta}})/\sigma^2_{\pi,j} - 1\}$. Thus, in this case one should choose $\tilde{\boldsymbol{\theta}}_0$ so that $\sigma^2_{\pi_0,j}$ is sufficiently close to $\bar{V}_j(\tilde{\boldsymbol{\theta}})$ from below to ensure this inequality. For example, if $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ is a beta distribution, $\mathrm{Be}(\tilde{\alpha}, \tilde{\beta})$, then σ^2_{π} $= \mu(1-\mu)/(\tilde{\alpha} + \tilde{\beta} + 1)$ and $\bar{V}(\tilde{\boldsymbol{\theta}}) = \mu(1-\mu)$, where $\mu = \tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$. One may construct $\pi_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ as a $\mathrm{Be}(\tilde{\alpha}/c, \tilde{\beta}/c)$ by choosing $c > 0$ large enough so that $1/c < \epsilon$, since this implies that the above inequality holds.

Alternatively, a proper non-informative prior could be considered for $\pi_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$. However, this might not be appropriate for defining ESS. For example, the Jeffreys prior for a binomial success probability is $\mathrm{Be}(\frac{1}{2}, \frac{1}{2})$, which gives ESS values that are too large by an additive factor of 1 due to the implied sample size of 1.

To see how the definition works for $\boldsymbol{\theta}$ of dimension > 1, consider a bivariate normal distribution with mean $\tilde{\mu} = (\tilde{\mu}_1, \tilde{\mu}_2)'$, variances $\tilde{\sigma}^2_1$, $\tilde{\sigma}^2_2$ and covariance $\tilde{\sigma}_{12}$. An ε-information prior is specified by using the same means $\tilde{\mu}$ but variances $c_1^2\tilde{\sigma}^2_1$ and $c_2^2\tilde{\sigma}^2_2$ and covariance $c_1 c_2 \tilde{\sigma}_{12}$ for arbitrarily large $c_1 > 0$ and $c_2 > 0$. As a practical guideline, we suggest choosing c large enough so that a further increase would not change the ESS by more than 0.1.

Given the likelihood $f(\mathbf{Y}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^n f(Y_i \mid \boldsymbol{\theta})$ of an i.i.d. sample $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ and ε-information prior $\pi_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$, the posterior is

$$\pi_n(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_n) \propto \pi_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) f(\mathbf{Y}_n \mid \boldsymbol{\theta}).$$

MTM define a distance between the prior $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ and $\pi_n(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_n)$ obtained from a hypothetical sample $\mathbf{Y}_n$ of size n starting with π_0. To do this, MTM use the prior variance $\sigma^2_{\pi,j}$ and the average posterior variance $\sigma^2_{\pi_0,j,n}$ under the ε-information prior. The average is with respect to the prior predictive distribution $f(\mathbf{Y}_n|\tilde{\boldsymbol{\theta}}) = \int f(\mathbf{Y}_n \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta}$. Let $\sigma^2_{\pi_0,j,n}(\mathbf{Y}_n)$ denote the posterior variance of θ_j conditional on $\mathbf{Y}_n$, under the ε-information prior π_0. Then $\sigma^2_{\pi_0,j,n} = \int \sigma^2_{\pi_0,j,n}(\mathbf{Y}_n)\ df(\mathbf{Y}_n|\tilde{\boldsymbol{\theta}})$. In cases where these variances cannot be computed analytically, we use approximations, denoted by $\widehat{\sigma}^{-2}_{\pi,j}$ and $\widehat{\sigma}^{-2}_{\pi_0,j,n}$. For example, one could use the negative second derivatives of the log distributions, as in MTM. We define the distance between π and $\pi_{0,n}$ to be

$$\Delta(n, \pi, \pi_0) \ = \ \sum_{j=1}^{d} |\sigma^{-2}_{\pi,j} - \sigma^{-2}_{\pi_0,j,n}|. \tag{9}$$

When interest focuses on only one of the parameters, say θ_j, then we use $|\sigma_{\pi,j}^{-2} - \sigma_{\pi_0,j,n}^{-2}|$ to define a distance between the marginal prior on θ_j versus the marginal posterior on θ_j under the ε-information prior and a sample of size n.

Definition of ESS in the non-CIHM case: *The effective sample size (ESS) of a parametric prior $\pi(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ for which $\sigma_{\pi,j}^2 < \infty$ for all $j = 1, ..., d$, with respect to the likelihood $f(\mathbf{Y}_n \mid \boldsymbol{\theta})$* is the interpolated integer n that minimizes the distance $\Delta(n, \pi, \pi_0)$.

Computation of $\Delta(n, \pi, \pi_0)$ is carried out either analytically or using a simulation-based numerical approximation.

3.2 Conditionally independent hierarchical models

We extend the definition of ESS to accommodate two-level CIHMs in a balanced study design with K subgroups each having sample size m, i.e., $\mathbf{Y}_k = (Y_{k,1}, \cdots, Y_{k,m})$ for each $k = 1, \cdots, K$ and the total sample size of $\mathcal{Y}_M = (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ is $M = m \times K$. To accommodate the hierarchial structure, we propose the following two alternative definitions of ESS for CIHMs. Recall the discussion following (1). Under a CIHM interest may focus on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ (case 1) or on $\tilde{\boldsymbol{\theta}}$ (case 2).

Case 1: When $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ are the parameters of primary interest, the ESS is a function of the target marginal prior $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ and the sampling model $f(\cdot)$. In this case, we constructively define an ε-information prior, $\pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, as follows. First, specify an ε-information prior $\pi_{1,0}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ of $\pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, as described in Section 3.1, and use this to define $\pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi}) = \int \pi_{1,0}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) \pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}) d\tilde{\boldsymbol{\theta}}$. A proof that $\pi_{12,0}$ defined in this way is in fact an ε-information prior is given in the appendix.

An alternative way to define an ε-information prior for case 1 could be to first specify $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ and then compute $\pi_{12,0}$. However, this approach is tractable only if $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ can be specified analytically. Consequently, we do not use this alternative approach.

Given the likelihood $f(\mathcal{Y}_M \mid \boldsymbol{\theta}) = \prod_{k=1}^K f(\mathbf{Y}_k \mid \boldsymbol{\theta}_k)$ and $\pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, we denote the hypothetical posterior by $\pi_{12,M}(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \mathcal{Y}_M) \propto \pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi}) f(\mathcal{Y}_M \mid \boldsymbol{\theta})$. Denote $\bar{\boldsymbol{\theta}}_{12} = \mathrm{E}_{\pi 12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, the prior mean of $\boldsymbol{\theta}$ under $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ and let $\overline{\mathcal{Y}}_M = E(\mathcal{Y}_M \mid \bar{\boldsymbol{\theta}}_{12})$ denote the prior mean of $\mathcal{Y}_M$ under $\bar{\boldsymbol{\theta}}_{12}$. We will later discuss how to proceed in problems where $\overline{\mathcal{Y}}_M$ is not meaningfully defined, due to censoring, the use of covariates, and other reasons. Let $\Sigma_{\pi_{12}}$ denote the marginal variance-covariance matrix of $\boldsymbol{\theta}$ under π_{12}, and similarly for $\Sigma_{\pi_{12,M}}(\overline{\mathcal{Y}}_M)$ under $\pi_{12,M}(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \overline{\mathcal{Y}}_M)$. Often exact evaluation is not possible and approximations $\widehat{\Sigma}_{\pi_{12}}$ and $\widehat{\Sigma}_{\pi_{12},M}$ must be used. For example, one could use the negative inverse Hessian matrices of $\log(\pi_{12})$ and $\log(\pi_{12,M})$ evaluated at $\bar{\boldsymbol{\theta}}_{12}$. Similarly to $\Delta(\cdot)$ in equation (9) for a non-hierarchical model, in case 1 of a CIHM we define the distance between $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ and $\pi_{12,M}(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \overline{\mathcal{Y}}_M)$ to be

$$\Delta_1(M, \pi_1, \pi_2, \pi_{1,0}) = \left| \det(\Sigma_{\pi_{12}}^{-1}) - \det\{\Sigma_{\pi_{12,M}}^{-1}(\overline{\mathcal{Y}}_M)\} \right|. \quad (10)$$

Definition of ESS for Case 1: *The ESS of* $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ *with respect to* $f(\cdot)$, denoted by $\text{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, is the interpolated value of the sample size M minimizing $\Delta_1(M, \pi_1, \pi_2, \pi_{1,0})$.

We use determinants in the definition of Δ_1 to incorporate the off-diagonal elements of the variance/covariance matrices and thereby account for association induced by the hyperprior among the parameters, $\theta_1, \ldots, \theta_K$, which is a key aspect of a CIHM. The determinants quantify the total amounts of variability of the prior π_{12} and the hypothetical posterior $\pi_{12,M}$. The same definition of distance could be used for non-hierarchical models in place of (9). However, the trace used in (9) is easier to evaluate and leads to exactly the same ESS in all cases that we considered. Although our choice seems to make sense intuitively, it still is arbitrary. We validate the definition by investigating important cases of CIHMs, given in Section 5 and the supplementary materials.

Case 2: When the hyperparameters $\tilde{\boldsymbol{\theta}}$ are of primary interest, the ESS is a function of the target prior $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ and the marginal likelihood (4). We define an ε-information prior, $\pi_{2,0}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0)$, and use the marginal likelihood $f_1(\mathcal{Y}_M \mid \tilde{\boldsymbol{\theta}})$ to update $\pi_{2,0}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0)$ to obtain the hypothetical posterior $\pi_{2,M}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0, \mathcal{Y}_M)$. Let $\bar{\tilde{\boldsymbol{\theta}}}_2 = \text{E}_{\pi_2}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ denote the prior mean of $\tilde{\boldsymbol{\theta}}$ under $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$, and let $\overline{\mathcal{Y}}_{M,2} = E(\mathcal{Y}_M \mid \bar{\tilde{\boldsymbol{\theta}}}_2)$. We define a distance between $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ and $\pi_{2,M}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0, \mathcal{Y}_M)$ for sample size M as in the definition in (11), using the variance/covariance matrices Σ_{π_2} under the prior $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ and $\Sigma_{\pi_{2,M}}(\overline{\mathcal{Y}}_{M,2})$ under $\pi_{2,M}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0, \overline{\mathcal{Y}}_{M,2})$. When the variance/covariance matrices are not available in closed form, one can again use a numerical approximation.

Definition of ESS for Case 2: *The ESS of* $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ *with respect to* $f_M(\mathcal{Y}_M \mid \tilde{\boldsymbol{\theta}})$, denoted by $\text{ESS}_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$, is the interpolated value of M minimizing

$$\Delta_2(M, \pi_2, \pi_{2,0}) = \left| \det(\Sigma_{\pi_2}^{-1}) - \det\{\Sigma_{\pi_{2,M}}^{-1}(\overline{\mathcal{Y}}_{M,2})\} \right|. \tag{11}$$

Definition (11) is equivalent to the non-hierarchical ESS definition, simply because after marginalizing with respect to the group specific parameters θ_k the hierarchical model (1) reduces to a non-hierarchical model. If interest is focused on a subvector $\tilde{\boldsymbol{\theta}}_s$ of $\tilde{\boldsymbol{\theta}}$, the ESS can be determined similarly in terms of the marginal hyperprior $\pi_2(\tilde{\boldsymbol{\theta}}_s \mid \boldsymbol{\phi})$.

An important aspect of the definition is that the ESS could be infinite when it is not possible to achieve a comparably informative posterior with any finite sample size. For example, recall the stylized normal/normal CIHM introduced after (1). Assume we fix $\tilde{\gamma}^2 = 1$, $\tilde{\mu} \sim N(0, 0.01)$ and $K = 10$. In particular, we keep the number K of subpopulations fixed and consider an increasing number of samples per group. No matter how many observations, it is impossible to match the prior variance $\text{Var}(\tilde{\mu}) = 0.01$ as the posterior variance under $\pi_{0,n}$. Formally, the minimization (11) has no solution. This is desireable. The informative prior $\tilde{\mu} \sim N(0, 0.01)$ cannot be interpreted as the posterior under any hypothetical sample having finite sample size, because, in this example, the number of groups is fixed at $K = 10$. However, while no sample size of a hypothetical prior study under the same experimental design, i.e., fixed K, can justify the employed prior, it still is possible to report a meaningful finite ESS under an alternative experimental design. For example, one could report the required sample size of a hypothetical prior study with fixed $m = 20$, $\sigma^2 = 1$, and increasing K. In this example we find ESS = 2100 with $K = 105$.

3.3 Algorithm for ESS Computation

Let $\bar{\boldsymbol{\theta}}_{12}$ denote the prior mean vector of $\boldsymbol{\theta}$ under $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, and let $\bar{\tilde{\boldsymbol{\theta}}}_2$ denote the prior mean vector of $\tilde{\boldsymbol{\theta}}$ under $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$. Let $M_{max} = m_{max} \times K$ be a positive integer multiple of K chosen so that it is reasonable to assume that $ESS \leq M_{max}$.

Algorithm 1, ESS for Case 1:

Step 1. Evaluate the target marginal prior $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$.

Step 2. Evaluate ε-information prior $\pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$.

Step 3. For each $M = 0, K, 2K, \ldots, M_{max}$, compute $\Delta_1(M, \pi_1, \pi_2, \pi_{1,0})$.

Step 4. $\text{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ is the interpolated value of M minimizing $\Delta_1(M, \pi_1, \pi_2, \pi_{1,0})$.

When $\Sigma_{\pi_{12}}^{-1}$ cannot be evaluated analytically, then one can use the negative Hessian of the log prior as approximation $\widehat{\Sigma}_{\pi_{12}}^{-1} = -H_{\pi_{12}}(\overline{\boldsymbol{\theta}}_{12})$. The Hessian $H_{\pi_{12}}$ is the matrix of second derivatives of the $\log(\pi_{12})$ with respect to θ_j. The Hessian is evaluated at $\overline{\boldsymbol{\theta}}_{12}$. When $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ itself is not available in closed form, then we use the following simulation-based approximation. First, simulate a Monte Carlo sample $\tilde{\boldsymbol{\theta}}^{(1)}, \ldots, \tilde{\boldsymbol{\theta}}^{(T)}$ from $\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ for large T, e.g. $T =$ 100,000. Use the Monte Carlo average $T^{-1} \sum_{t=1}^{T} \frac{\partial}{\partial \theta_k} \pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^{(t)})$ in place of $\int \frac{\partial}{\partial \theta_k} \pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) \pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}) d\tilde{\boldsymbol{\theta}}$, and similarly for the second order partial derivatives. When $\bar{\boldsymbol{\theta}}_{12} = \text{E}_{\pi_{12}}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ cannot be computed analytically, first simulate a Monte Carlo sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ from $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}) = \int \pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) \pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}) d\tilde{\boldsymbol{\theta}}$ for large T, and compute the mean $T^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}$. Similarly, when $\Sigma_{12,M}^{-1}(\overline{\mathcal{Y}}_M)$ is not available in closed form one can use the negative Hessian of $\log \pi_{12,M}(\overline{\boldsymbol{\theta}}_{12} \mid \overline{\mathcal{Y}}_M, \boldsymbol{\phi})$ as a convenient approximation, $\widehat{\Sigma}_{12,M}^{-1}(\overline{\mathcal{Y}}_M) = -H_{\pi_{12,M}}(\overline{\boldsymbol{\theta}}_{12}, \overline{\mathcal{Y}}_M)$. The posterior $\pi_{12,M}$ is evaluated conditional on $\overline{\mathcal{Y}}_M$ and the Hessian is evaluated at $\overline{\boldsymbol{\theta}}_{12}$.

Algorithm 2, ESS for Case 2:

Step 1. Specify $\pi_{2,0}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}_0)$.

Step 2. For each $M = 0, K, 2K, \ldots, M_{max}$, compute $\Delta_2(M, \pi_2, \pi_{2,0})$.

Step 3. $\text{ESS}_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ is the interpolated value of M minimizing $\Delta_2(M, \pi_2, \pi_{2,0})$.

When $\Sigma_{\pi_2}^{-1}$ or $\Sigma_{\pi_2,M}^{-1}(\overline{\mathcal{Y}}_{M,2})$ is not available in closed form we proceed similarly as before, using the negative Hessians to approximate $\Sigma_{\pi_2}^{-1}$ or $\Sigma_{\pi_2,M}^{-1}$. The posterior $\pi_{2,M}$ is evaluated conditional on $\overline{\mathcal{Y}}_{M,2}$. The Hessian is evaluated at $\bar{\tilde{\boldsymbol{\theta}}}_2$. In many cases, steps 2 and 3 may be repeated to compute $\text{ESS}_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})$ for subvectors of $\tilde{\boldsymbol{\theta}}$ that are of interest.

3.4 Numerical Evaluation of ESS

Equations (10) and (11) define a distance between prior and posterior as a difference of determinants of a prior precision matrix versus a posterior precision matrix. The posterior precision matrix is evaluated using the plug-in values $\overline{\mathcal{Y}}_M$ in Δ_1 and $\overline{\mathcal{Y}}_{M,2}$ in Δ_2, respectively. In some studies, however, the definition of $\overline{\mathcal{Y}}_M$ or $\overline{\mathcal{Y}}_{M,2}$ as an expectation is simply not meaningful. This is explained most easily by the following examples. Many phase I studies involve choosing doses for future patients as a function of the outcomes of earlier patients. In such settings, it is not meaningful to represent a sample of size M by a mean $\overline{\mathcal{Y}}_M$ or $\overline{\mathcal{Y}}_{M,2}$. More generally, *outcome adaptive designs* for clinical studies make it difficult to meaningfully define $\overline{\mathcal{Y}}_M$. Similar difficulties arise when studies involve censored outcomes or regression on covariates.

In these cases, we use an extended definition of Δ_1 that replaces the plug in of $\overline{\mathcal{Y}}_M$ by full prior predictive marginalization. Specifically, we replace $\det\{|\Sigma_{\pi_{12},M}^{-1}(\overline{\mathcal{Y}}_M)|\}$ by an expectation of $\det\{|\Sigma_{\pi_{12},M}^{-1}(\mathcal{Y}_M)|\}$, averaging over possible study realizations $\mathcal{Y}_M$ with M patients. The sampling scheme might include additional variables, like assigned dose, patient-specific covariates, or censoring times. Let x_i denote these additional variables for the i^{th} patient. Let $g(x_i \mid Y_1, \ldots, Y_{i-1}, x_1, \ldots, x_{i-1}, \overline{\boldsymbol{\theta}}_{12})$ denote a probability model for x_i. When x_i is a deterministic function of the conditioning variables, we interpret $g(\cdot)$ as a degenerate distribution with a single point mass. For example, in the case of an adaptive dose escalation design, the dose for the i^{th} patient is a function of the earlier outcomes. Thus, the outcomes are not independent, making the interpretation of $\overline{\mathcal{Y}}_M$ tricky. We replace use of the plug in $\overline{\mathcal{Y}}_M$ by the expectation of $\det\{|\Sigma_{\pi_{12},M}^{-1}(\mathcal{Y}_M)|\}$ with respect to the marginal model,

$$f_M(\mathcal{Y}_M \mid \overline{\boldsymbol{\theta}}_{12}) = \int \prod_{i=1}^{M} f(Y_i \mid Y_1, \ldots, Y_{i-1}, x_i, \overline{\boldsymbol{\theta}}_{12}) \\ g(x_i \mid Y_1, \ldots, Y_{i-1}, x_1, \ldots, x_{i-1}, \overline{\boldsymbol{\theta}}_{12})\, \mathrm{d}x_1 \cdots \mathrm{d}x_M. \quad (12)$$

In words, $f_M(\cdot)$ is the marginal distribution of possible study outcomes when we carry out the study under a given design, and the prior mean $\overline{\boldsymbol{\theta}}_{12}$. The predictive model (12) highlights the fact that ESS is defined in the context of an assumed experimental design. This includes details like the expected extent of censoring, distribution of covariates x_i, and more. Recall the stylized normal/normal CIHM stated below (1). One could consider two alternative experiments, either increasing the number of observations within each subpopulation, $k = 1, \ldots, K$ for fixed K, or alternatively one could consider increasing the number of groups K with a fixed number of observations per group. The two experiments naturally report different ESS values for the same prior, simply because the information that is being accrued with hypothetical future patients differs.

In the definition of the ESS, Case 1, for such studies we replace $\det\{|\Sigma^{-1}_{\pi_{12,M}}(\overline{\mathcal{Y}}_M)|\}$ in the definition of Δ_1 by

$$\int \det\{\Sigma^{-1}_{\pi_{12,M}}(\mathcal{Y}_M)\}\; f_M(\mathcal{Y}_M \mid \overline{\boldsymbol{\theta}}_{12}) d\mathcal{Y}_M \tag{13}$$

where $\Sigma^{-1}(Y_M) = \mathrm{Var}(\boldsymbol{\theta} \mid \mathcal{Y}_M)$ is the posterior variance under data $\mathcal{Y}_M$ and prior π_{12}. Similarly, we replace $\det\{|\Sigma^{-1}_{\pi_{2,M}}(\overline{\mathcal{Y}}_{M,2})|\}$ in the definition of Δ_2 in Case 2 by

$$\int \det\{\Sigma^{-1}_{\pi_{2,M}}(\mathcal{Y}_M)\}\; f_{M,2}(\mathcal{Y}_M \mid \overline{\tilde{\boldsymbol{\theta}}}_2) d\mathcal{Y}_M \tag{14}$$

where

$$\begin{aligned} f_{M,2}(\mathcal{Y}_M \mid \overline{\tilde{\boldsymbol{\theta}}}_2) = \int\!\!\int \prod_{i=1}^{M} & f(Y_i \mid Y_1, \ldots, Y_{i-1}, x_i, \boldsymbol{\theta}) \\ & \times g(x_i \mid Y_1, \ldots, Y_{i-1}, x_i, \ldots, x_{i-1}, \boldsymbol{\theta})\, \mathrm{d}x_1 \cdots \mathrm{d}x_M\ \pi_1(\boldsymbol{\theta} \mid \overline{\tilde{\boldsymbol{\theta}}}_2)\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{15}$$

For the standard CIHMs discussed later, in Section 5, the definitions using (13) instead of $\det\{|\Sigma^{-1}_{\pi_{12,M}}(\overline{\mathcal{Y}}_M)|\}$ and (14) in place of $\det\{|\Sigma^{-1}_{\pi_{2,M}}(\overline{\mathcal{Y}}_{M,2})|\}$ in the definitions of Δ_1 and Δ_2, respectively, give the same ESS as the original definitions. That is why we recommend using the simplified definitions (10) and (11) when possible. We recommend using the extended definitions only for problems where the definition of $\overline{\mathcal{Y}}_M$ or $\overline{\mathcal{Y}}_{M,2}$ is not meaningful.

Algorithms 1 and 2 remain unchanged. The only additional detail is the evaluation of Δ_1 and Δ_2. For Δ_1, we use numerical evaluation of the integral (13).

Algorithm 3, Numerical Evaluation of Δ_1: Evaluation of (13) as a Monte Carlo integral.

```
Repeat $L$ iterations of Steps 1 – 6:
    Step 1: Generate $Y_1 \sim f(Y_1 \mid x_1, \overline{\boldsymbol{\theta}}_{12})$.
    Repeat for $i = 2, \ldots, M$:
        Step 2: Evaluate $g(x_i \mid Y_1, \ldots, Y_{i-1}, x_i, \ldots, x_{i-1}, \overline{\boldsymbol{\theta}}_{12})$. If $g(\cdot)$ is a deterministic function (as in dose allocation), record the value as $x_i$. If $g(\cdot)$ is a non-degenerate distribution generate $x_i$ accordingly.
```

Step 3: Generate $Y_i \sim f(Y_i \mid Y_1, \ldots, Y_{i-1}, x_i, \overline{\boldsymbol{\theta}}_{12})$, using x_i from Step 2.

Step 4: Simulate the outcome vector $\mathcal{Y}_M = (Y_1, \ldots, Y_M)$.

Step 5: Evaluate the posterior variance covariance matrix $\Sigma(\boldsymbol{\theta} \mid \mathcal{Y}_M)$. This evaluation might require Markov chain Monte Carlo integration when the model is not conjugate.

Step 6: Evaluate $\det(\Sigma^{-1}(\mathcal{Y}_M))$.

Step 7: Approximate the desired integral by averaging over all L iterations of Steps 1 through 6,

$$\frac{1}{L}\sum \det(\Sigma^{-1}(\mathcal{Y}_M)) \approx \int \det\{\Sigma^{-1}(\mathcal{Y}_M)\}\, f_M(\mathcal{Y}_M \mid \overline{\boldsymbol{\theta}}_{12}) dY_M, \tag{16}$$

The sum is over the L repeat simulations of Steps 1 through 6, plugging in the vector $\mathcal{Y}_M$ that is generated in each step of the simulation.

With a minor variation, the same algorithm can be used to evaluate (14). Step 1 is replaced by

Step 1′: Generate $\boldsymbol{\theta} \sim \pi_1(\boldsymbol{\theta} \mid \overline{\tilde{\boldsymbol{\theta}}}_2)$ and generate $Y_1 \sim f(Y_1 \mid x_1, \boldsymbol{\theta})$.

The rest of the algorithm for computing ESS remains unchanged, with $\boldsymbol{\theta}$ replacing $\overline{\boldsymbol{\theta}}_{12}$. It is important that a new value $\boldsymbol{\theta}$ is generated for each of the L repetitions of Steps 1 through 6. This approximates the outer integral with respect to $\boldsymbol{\theta}$.

4 Application to the motivating examples

4.1 ESS for the sarcoma trial

The goal of the sarcoma trial was to estimate the efficacy of a molecularly targeted drug within each disease subtype. Recall that ξ_k was the response probability in subtype k, and $\theta_k = \log\{\xi_k/(1-\xi_k)\}$. The likelihood for M patients with m patients in each subtype having response indicators $\mathbf{Y}_{k,m} = (Y_{k1}, \ldots, Y_{km})$ is

$$f(\mathcal{Y}_M \mid \boldsymbol{\theta}) \;\propto \prod_{k=1}^{K}\prod_{i=1}^{m} \xi_k(\theta_k)^{Y_{ki}}\{1-\xi_k(\theta_k)\}^{1-Y_{ki}}. \tag{17}$$

For this example, based on the normal-gamma hierarchical prior described in Section 2, a closed form of the target prior in case 1 can be found analytically, as we will show in Section 5. Based on this prior and the likelihood (17), applying Algorithm 1 of subsection 3.3 gives $\mathrm{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}} = (-1.386, 10), \boldsymbol{\phi}_{\tilde{\gamma}} = (2, 20)) = 2.6$. This ESS value is reasonable, since a separate early stopping rule was applied for each subtype in this study. For comparison, if one assumes instead that $\boldsymbol{\phi}_{\tilde{\mu}}$ = (-1.386,1) and $\boldsymbol{\phi}_{\tilde{\gamma}} = (20, 1)$, then $\mathrm{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}}) = 702.2$. This illustrates the important practical point that a seemingly reasonable choice of $\boldsymbol{\phi}_{\tilde{\mu}} = (\mu_\phi, \tau^2_\phi)$ and $\boldsymbol{\phi}_{\tilde{\gamma}} = (a_\phi, b_\phi)$ may give an excessively informative prior. Figure 1 gives plots of the ESS_{12} values as a function of τ_ϕ^2 for each of the four pairs $\boldsymbol{\phi}_{\tilde{\gamma}}$ = (2,20), (5,20), (10,20), and (10,10), when μ_ϕ = -1.386. Figure 1 shows that, as the variance parameter τ_ϕ^2 in the hyperprior of $\tilde{\mu}$ gets larger, the ESS_{12} values decrease, which is intuitively correct. Similarly, ESS_{12} increases with the prior mean of the precision parameter $\tilde{\gamma}$ under $\pi_2(\tilde{\gamma} \mid \boldsymbol{\phi}_{\tilde{\gamma}})$.

If one wishes to focus, instead, on the overall mean treatment effect across all disease subtypes, (case 2), it leads to consideration of the hyperprior $\pi_2(\tilde{\mu} \mid \boldsymbol{\phi}_{\tilde{\mu}})$. The ESS in this case may be computed using Algorithm 2 of subsection 3.3, which gives $\mathrm{ESS}_2(\tilde{\mu} \mid \boldsymbol{\phi})$ = 3.7. We conclude that the hierarchical prior used by Thall et al. (2003) is reasonable under both points of view, either when one is concerned with inference on success probabilities for the sarcoma subtypes, or when one is interested in the average response probability for all sarcomas.

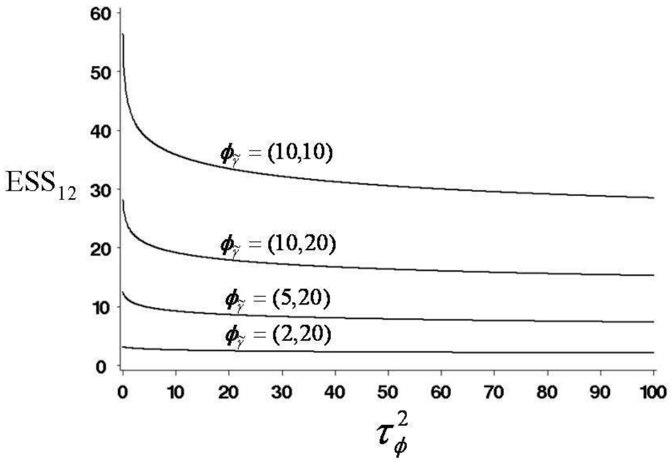

Figure 1: Plots of ESS_{12} as a function of variance parameter τ_ϕ^2 in the hyperprior of $\tilde{\mu}$ for the four sets of $\boldsymbol{\phi}_{\tilde{\gamma}}$ = (2,20), (5,20), (10,20), and (10,10), when μ_ϕ = -1.386.

4.2 A CRM Dose-finding Trial for Multiple Patient Subgroups

In the second example, the goal of the CRM-type dose-finding study is to find the optimal dose in each subgroup. Recall that the probability of toxicity in each subgroup is modeled by $p_k(x_i, \boldsymbol{\theta}_k)$ with $\mathrm{logit}\{p_k(x_i, \boldsymbol{\theta}_k)\} = \alpha_k + \beta_k x_i$, where $\boldsymbol{\theta}_k = (\alpha_k, \beta_k)$. Let m_1, m_2, m_3, m_4 denote the number of patients in the four subgroups, let $x_{k,[i]}$ denote the dose assignment for the i^{th} patient in subgroup k, and let $D_{m_k} = \{Y_{k,1}, x_{k,[1]}, \ldots, Y_{k,m_k},$

$x_{k,[m_k]}\}$ denote the dose assignments and the responses observed in subgroup k. The likelihood for all $M = \sum_{k=1}^{4} m_k$ patients is

$$f(\mathcal{Y}_M \mid \mathbf{x}_M, \boldsymbol{\theta}) \;\propto \prod_{k=1}^{4} \prod_{i=1}^{m_k} p_k(x_{k,[i]}, \boldsymbol{\theta}_k)^{Y_{k,i}} \{1 - p_k(x_{k,[i]}, \boldsymbol{\theta}_k)\}^{1-Y_{k,i}}. \tag{18}$$

Following the extended ESS definition given in Section 3.4, we compute ESS values under several combinations of $(\sigma^2_{\alpha,\phi}, \sigma^2_{\beta,\phi}, U_\phi)$ for $\sigma^2_{\alpha,\phi} = \sigma^2_{\beta,\phi}$ using Algorithm 3. Dose levels are determined on the basis of the observed data. As mentioned before, ESS is always defined in the context of an assumed experimental design. In the case of this study, the assumed relative accrual rates $(0.4, 0.3, 0.2, 0.1)$ are part of the design, and are used in the evaluation of ESS.

Posterior summaries are evaluated using Markov chain Monte Carlo posterior simulation (with three parallel chains of length 5,000 with a burn-in of 500). The posterior means at the six dose levels in each subgroup are required in the dose-escalation/de-escalation algorithm for the respective new cohort. Table 1 summarizes the results of the computations, including the subgroup ESS values based on the assumed relative accrual rates $(0.4, 0.3, 0.2, 0.1)$. Those computations suggest that, considering the maximum sample size of 36, the prior choices that are summarized in the first two rows of the table are defensible. The implied ESS is reasonably small compared to the sample size, since in these cases the per-subgroup sample size is less than 2. The prior in the third row may or may not be considered suitably non-informative for a dose-finding trial, but the prior in row 4 clearly is far too informative. Without computing ESS, it would be very difficult to make such determinations.

Table 1: ESS values computed for several sets of level 2 prior parameters $(\sigma^2_{\alpha,\phi}, \sigma^2_{\beta,\phi}, U_\phi)$. The per-subgroup ESS values are based on assumed relative accrual rates .40, .30, .20, .10 for k = 1, 2, 3, 4.

			Per-subgroup ESS			
$\sigma^2_{\alpha,\phi} = \sigma^2_{\beta,\phi}$	U_ϕ	ESS	1	2	3	4
100	5	6.6	2.6	2.0	1.3	0.7
25	5	7.1	2.8	2.1	1.4	0.7
4	5	9.1	3.6	2.7	1.8	0.9
25	2	25.8	10.3	7.7	5.2	2.6

Computation of frequentist operating characteristics (OCs) by computer simulation, and calibration of prior hyperparameters as well as design parameters on that basis, has become standard practice in Bayesian clinical trial design. An important question is how computing the prior's ESS can improve this process. This is especially important logistically because obtaining both prior ESS and the design's OCs are computationally intensive. However, computing an ESS takes even less time than the time required for OCs under only one scenario. This is the case because in contrast to OC simulation, the evaluation of ESS does not require the simulation of complete trial histories. Also, in practice, after establishing hyperparameters that determine prior mean values, one guesses numerical values of second order hyperparameters, including variances and correlations, simulates the trial using that prior, and examines the resulting OCs. Since the numerical values of prior hyper-variances have no meaning *per se*, in the absence of ESS, one must rely on the OCs alone to determine whether the prior is acceptable and reasonable. This process often is repeated many times, until satisfactory choices are found, and is quite time-consuming and tedious. When the ESS of each prior is computed, however, it may be used to calibrate the hyper-variances so that they give a reasonable ESS, before running any simulations of complete trial histories. This saves a great deal of time because, as noted above, computing an ESS is far less time-consuming than running simulations.

Table 2: Simulation study of the CRM dose-finding trial with heterogeneous subgroups assuming the prior with level 2 prior parameters $\sigma^2_{\alpha,\phi} = \sigma^2_{\beta,\phi} = 25$ and $U_\phi = 2$. The percentage of times that the method selected each dose level as the final MTD in each subgroup (%recommendation) and the numbers of patients who were treated at each dose level (No. of patients) out of 36 patients are summarized. Correct selections are marked in bold-face.

Subgroup	Dose level	d_1	d_2	d_3	d_4	d_5	d_6
1	True prob. tox	.05	.10	.15	.30	.50	.65
	%recommendation	.05	.16	.30	**.36**	.07	.04
	No. of patients	2.1	3.1	4.5	**3.6**	1.1	.4
2	True prob. tox	.10	.20	.30	.45	.60	.70
	%recommendation	.08	.26	**.39**	.24	.02	.01
	No. of patients	1.8	2.9	**3.3**	2.0	.6	.1
3	True prob. tox	.15	.30	.45	.55	.65	.75
	%recommendation	.12	**.34**	.32	.16	.02	.01
	No. of patients	1.2	**2.2**	1.9	1.0	.3	.1
4	True prob. tox	.10	.10	.10	.20	.20	.20
	%recommendation	.05	.20	.32	.31	.09	.02
	No. of patients	.5	.9	1.0	.8	.2	.1

Aside from obtaining a design's OCs, because the ESS is readily interpretable, it is a useful tool for deciding whether a prior is reasonable *per se*. Recall that the ESS is a property of the prior, likelihood, and experimental design. In contrast, obtaining the OCs of a given design also requires the specification of particular scenarios (fixed outcome probabilities) under which the design is evaluated. In the absence of ESS, it is possible to obtain OCs that look reasonable when in fact the prior is undesirably informative. That is, one may unknowingly be misled by simulation results. This is especially worrying when the clinical outcome is a complex function of treatment, as in a dose-response setting. As an example, consider the CRM dose-finding trial with heterogeneous subgroups under the prior with level 2 prior parameters $\sigma^2_{\alpha,\phi} = \sigma^2_{\beta,\phi} = 25$ and $U_\phi = 2$ in the last row of Table 1. Recall that the ESS = 25.8 for that prior. A simulation of the trial assuming these level 2 hyper-prior parameters is summarized in Table 2, which shows that, despite the overly informative prior in terms of ESS, the subgroup-specific CRM design is most likely to choose the right doses in subgroups 1, 2, and 3. In contrast, the dose selection is clearly wrong in subgroup 4, since the highest dose level would be most desirable in that subgroup. In the assumed scenario, the dose-toxicity curve in subgroup 4 is qualitatively different from the other subgroups, with true $p_4(x_i) = .10$ or .20 for all doses, well below the target of .30. Since subgroup 4 has the lowest accrual rate of 10% and thus accrues on average $.10 \times 36 = 3.6$ patients, the data from the trial cannot overcome the overly informative prior (ESS = 2.7) in this subgroup. Based on the ESS, however, one would never assume the prior with $\sigma^2_\phi = 25$ and $U_\phi = 2$. This example shows how use of the prior ESS complements the evaluation of OCs. In summary, in the context of Bayesian clinical trial design, prior ESS is both a tool for calibrating hyper-variances and thus speeding up the process of simulation to obtain the design's OCs, and also a simple summary statistic that helps investigators decide whether a prior is acceptable.

4.3 ESS for a Multicenter Randomized Trial

In the third example, recall that θ_{1k} and θ_{2k} denote the affective disorder recurrence rates in the two treatment arms for trial center k, and these are reparameterized as $\zeta_k = \log(\theta_{1k}/\theta_{2k})$ and $\eta_k = \log(\theta_{2k})$. Recall that $T_{jk,i}$ is the time to the event for patient i receiving treatment j at center k. Denote the observed time to an event at T or right-censoring by T^o and $\delta = I[T^o = T]$. Let $\mathcal{T}^o_M$ and $\boldsymbol{\delta}_M$ denote the vectors of all event observation times and indicators in the sample of $M = m \times K$ patients. The likelihood under the exponential model is

$$f(\mathcal{T}^o_M, \boldsymbol{\delta}_M \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{j=1}^{2} \prod_{i=1}^{m_{j,k}} \{\theta_{jk}\exp(-\theta_{jk}T_{jk,i})\}^{\delta_{jki}} \left\{\exp(-\theta_{jk}T^o_{jk,i})\right\}^{1-\delta_{jki}}, \tag{19}$$

where $\theta_{1k} = \exp(\zeta_k + \eta_k)$ and $\theta_{2k} = \exp(\eta_k)$. Recall that the hyperparameter $\tilde{\sigma}^2_\zeta$ characterizing inter-center variability is of primary interest (Case 2). Based on the hierarchical model described in Section 2 and the likelihood (19), we use Algorithm 2 to compute $\text{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi, s^2_\phi)$, which are $\text{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi = -1.61, s^2_\phi = 0.5^2) = 27.6$ and $\text{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi = 0, s^2_\phi = 0.5^2) = 9.1$. The two ESS values indicate that the first prior may be too informative, since the total sample size of the trial is 150. One may obtain a less informative hyperprior for $\tilde{\sigma}^2_\zeta$ by specifying larger s^2_ϕ. For example, $s^2_\phi = 0.75^2$, 1.0^2, and 1.25^2 give $\text{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi = 0, s^2_\phi) = 5.0$, 1.4, and 0.4. Table 3 summarizes the $\text{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi, s^2_\phi)$ values computed for several sets of (m_ϕ, s^2_ϕ) pairs.

As we mentioned in Section 2.3, the ESS computation in this example accounts for censoring in the process of numerical computation of the distance between the prior and posterior. For the above computations, we set-up the study duration at 2 years, as it was done in Stangl's study design. That is, patients are followed for 2 years, or until they experience the event. If patients do not experience the event by the end of the study, they are treated as censored cases. Given the event rates in the two treatment groups, the proportion of censored cases goes up as the study duration becomes shorter. When the study duration is 2 years, the ESS is computed to be 27.6, as shown in the first row of Table 3. When the study duration is 10 years, the ESS value is computed to be 24.5, while the ESS is 36.3 when the duration is 0.5 years. That is, as the proportion of censored cases rises, the ESS goes up, and vice-versa.

Table 3: $\mathrm{ESS}_2(\tilde{\sigma}^2_\zeta \mid m_\phi, s^2_\phi)$ values computed for several sets of (m_ϕ, s^2_ϕ), where $\tilde{\sigma}^2_\zeta$ represents the inter-center heterogeneity of the treatment effect, and m_ϕ and s^2_ϕ respectively denote the mean and variance parameters of the lognormal hyperprior of $\tilde{\sigma}_\zeta$.

m_ϕ	s^2_ϕ	ESS_2
-1.61	0.5^2	27.6
-1.61	0.75^2	12.1
-1.61	1.0^2	8.4
-1.61	1.25^2	6.6
-0.5	0.5^2	15.5
-0.5	0.75^2	8.0
-0.5	1.0^2	5.7
-0.5	1.25^2	2.1
0.0	0.5^2	9.1
0.0	0.75^2	5.0
0.0	1.0^2	1.4
0.0	1.25^2	0.4

From Figure 7 of Stangl (1996), it can readily be seen by inspection that prior 1 is too informative by comparing the posterior and prior of the 3rd stage variances. The ESS formalizes this judgment by quantifying the informativeness of prior 1. Moreover, the ESS validates prior 2 as being reasonable, whereas this in not obvious using only a graphical evaluation.

5 Validation in Some Standard CIHMs

The aim of the proposed ESS definitions is to provide an easily interpretable way to quantify the prior informativeness of commonly used CIHMs. The definition of ESS obtained by matching a given prior with a posterior distribution under an ε-information prior could be argued to be a natural choice. In order to formalize this, however, many detailed technical choices must be made. The definitions that we have proposed here are two of many possible formalizations. We arrived at the definitions after considering many alternatives, and evaluating the performance of each for several widely used CIHMs. The examples (one is in this section and two are in the supplementary materials) report the implied ESS for these models. As standard CIHMs, we use hierarchical extensions of standard models listed in Chapter 5 of Congdon (2005). In this section, we show how to compute the ESSs in cases 1 and 2 analytically in the normal CIHM with known sampling variance, and we argue that the obtained ESSs are sensible. In the supplementary materials, we provide computational details for a hierarchical model for binary responses recorded over different subpopulations as in the sarcoma trial of Section 2.1, and also for an alternative hierarchical model for the multicenter randomized trial analysis given in Section 2.3, assuming an exponential sampling model for the observed recurrence times.

Perhaps the most widely used CIHM is a hierarchical model with normal sampling model and conjugate prior and hyperprior. A typical example of such a CIHM is discussed in Gelman et al. (1995, Section 5.5), for a study of special coaching programs to prepare students for the Scholastic Aptitude Test (SAT). The observed outcomes are SAT scores, where Y_{ki} denotes score of the i-th student in school k. The scores are assumed to be normally distributed with school specific means $\theta_1, ..., \theta_K$ and known variance σ^2. The model is completed with conjugate hyperpriors, as follows:

$$\begin{array}{llcl}
\text{Sampling model} & Y_{k,1}, \cdots Y_{k,m} \mid \theta_k, \sigma^2 & \sim & \mathrm{N}(\theta_k, \sigma^2) \text{ indep. for all } k \\
\text{Prior} & \theta_k \mid \tilde{\mu}, \tilde{\gamma}^2 & \sim & \mathrm{N}(\tilde{\mu}, \tilde{\gamma}^2) \text{ i.i.d. for all } k \\
\text{Hyperpriors} & \tilde{\mu} \mid \mu_\phi, \tau_\phi^2 & \sim & \mathrm{N}(\mu_\phi, \tau_\phi^2) \\
 & \tilde{\gamma}^2 \mid \nu_\phi, S_\phi & \sim & \mathrm{Inv}-\chi^2(\nu_\phi, S_\phi).
\end{array}$$

Thus, for K schools, $\tilde{\mu}$ represents an overall effect and $\tilde{\gamma}^2$ represents inter-school variability. The fixed hyperparameters are $\boldsymbol{\phi}_{\tilde{\mu}} = (\mu_\phi, \tau_\phi^2)$ and $\boldsymbol{\phi}_{\tilde{\gamma}^2} = (\nu_\phi, S_\phi)$. Let $r = \tau_\phi^2/(\tilde{\gamma}^2 + \tau_\phi^2)$ denote the intra-class correlation. In this model the ESS in both cases 1 and 2 can be found analytically, because both the marginal prior π_{12} and the posterior $\pi_{12,M}$ are multivariate normal, thus the information matrices of the prior and posterior can be obtained analytically as the inverses of their variance-covariance matrices. Details are given in the supplementary materials.

Case 1: When the main goal is inference on the school-specific SAT scores $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, assuming fixed prior variance $\tilde{\gamma}^2$,

$$\mathrm{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}}) = \frac{\sigma^2}{\tilde{\gamma}^2} \left(\frac{1-r}{1+(K-1)r} \right)^{1/K} \times K. \tag{20}$$

This formula confirms what one would expect intuitively. The precision $1/\tilde{\gamma}^2$ of the conjugate normal prior for each school is equivalent to $\sigma^2/\tilde{\gamma}^2$ observations (SAT scores). Thus K independent priors for $\theta_1, ..., \theta_K$, corresponding to $r = 0$, would be equivalent to $K\sigma^2/\tilde{\gamma}^2$ observations. The second factor in (20) is obtained analytically from the determinant of the information matrix of π_{12} for the normal CIHM with conjugate priors (computational details are given in the supplementary materials). This factor accounts for the dependence among $\theta_1, \ldots, \theta_K$ in the two-level prior. Given K, as the correlation among $\theta_1, \ldots, \theta_K$ increases, this factor reduces the ESS_{12}. That is, as $r \to 1$, the $ESS_{12} \to 0$. In particular, for $K = 2$ schools, $\text{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}}) = \frac{\sigma^2}{\tilde{\gamma}^2}\left(\frac{1-r}{1+r}\right)^{1/2} \times 2$. Figure 2 shows a plot of $\text{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}})$ as a function of r and $\tilde{\gamma}^2$ for $K = 5$ and $\sigma^2 = 1.0$.

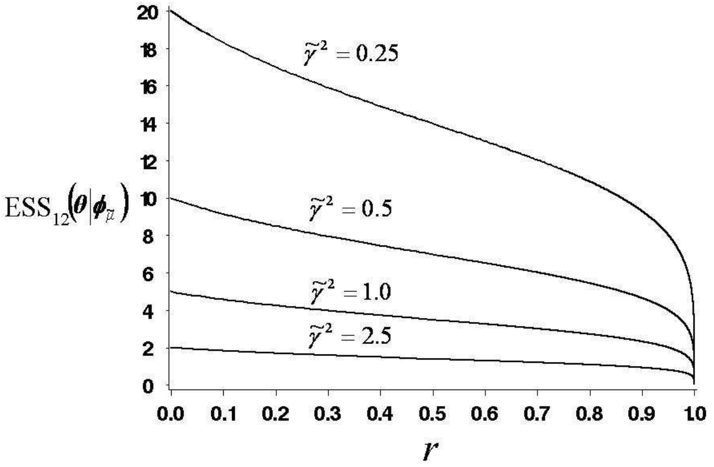

Figure 2: Plots of $\text{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}})$ as a function of prior variance $\tilde{\gamma}^2$ and r for the normal/inverse χ^2-normal-normal model, when $K = 5$ and $\sigma^2 = 1.0$.

To validate the plausibility of (20), we consider three limiting cases, each of which leads to a non-hierarchical model having an obvious prior ESS. (i) Given $\tilde{\gamma}^2$, as $r \to 0$, i.e. as $\tau_\phi^2 \to 0$, which is the limiting case with K independent schools, $\text{ESS}_{12} \to K \times \sigma^2/\tilde{\gamma}^2$ and $\sigma^2/\tilde{\gamma}^2$ is the ratio of the data variance to the prior variance, which is the prior ESS of a non-hierarchical model (Spiegelhalter et al. 1994, Section 3.1.2; MTM, Section 5, Example 3). (ii) In contrast, given $\tilde{\gamma}^2$, as $\tau_\phi^2 \to +\infty$, and consequently $r \to 1$, and there is one common effect $\theta = \theta_1 = \ldots = \theta_K$, i.e. this limiting case is a non-hierarchical model where schools are ignored. The marginal prior variance $\tilde{\gamma}^2 + \tau_\phi^2$ for the common effect diverges, however, which is reflected by the fact that $\text{ESS}_{12} \to 0$, as desired. (iii) On the other hand, consider the case where τ_ϕ^2 is fixed and $\tilde{\gamma}^2 \to 0$. Again, the model reduces to one common effect $\theta = \theta_1 = \ldots = \theta_K$, but now with finite marginal prior variance τ_ϕ^2. In this case, to learn with certainty that all effects are equal would require extremely reliable data, i.e., a very large number of students in each school, which is (correctly) reflected by the fact that $\mathrm{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_{\tilde{\mu}}) \to +\infty$. If we allow $\tilde{\gamma}^2$ to be random, $\mathrm{ESS}_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ takes the same form as expression (20), but with $\tilde{\gamma}^2$ replaced by its prior mean $\mu_{\tilde{\gamma}^2}$. (See Supplementary Materials, Section 1.)

Case 2: If the primary focus is the population parameter $\tilde{\mu}$, averaging across all schools,

$$\mathrm{ESS}_2(\tilde{\mu} \mid \boldsymbol{\phi}) = \frac{\sigma^2}{\tau_\phi^2}\left(1 - \frac{\mu_{\tilde{\gamma}^2}}{\tau_\phi^2 K}\right)^{-1}, \tag{21}$$

under the restriction that $\tau_\phi^2 > K^{-1}\mu_{\tilde{\gamma}^2}$. An explanation of this restriction is given in the supplementary materials, Section 1 (Normal/Inverse χ^2-Normal-Normal Model).

6 Discussion

We have proposed two practical definitions for the prior ESS of a conditionally independent hierarchical model. The main limitations are the need for numerical evaluations in many cases and the use of several arbitrary choices in the definitions. The arbitrary choices in the definition include the construction of an ε-information prior, evaluation of the posterior variance/covariance matrices conditional on $\overline{\mathcal{Y}}_M$ and $\overline{\mathcal{Y}}_{M,2}$, and evaluation of the distances Δ_1 and Δ_2 at the prior means $\bar{\boldsymbol{\theta}}_{12}$ and $\bar{\bar{\boldsymbol{\theta}}}_2$. The definitions of the two distances, based on the determinants of the information matrices, are reasonable but arbitrary. Alternative choices could be investigated. For example, one could use L_2 distance after a transformation of the variance-covariance matrix to some suitable standard form. For instance, the Cayley transformation maps the variance-covariance matrix Σ to the orthogonal matrix $(I-\Sigma)(I+\Sigma)^{-1}$. The main strengths of the method are the constructive nature of the definitions, and validation by matching prior ESS values obtained by the method with intuitively correct prior ESS values in special but important cases.

One of the reasons for the big success of hierarchical models, in particular, in inference for biomedical studies is the following feature. The hierarchical model constructs an informative prior for any given subgroup by borrowing strength from other subgroups and using the information contained in the data. If this feature is important, then the investigator might wish assurance that the hyperprior should not be excessively informative. The proposed ESS facilitates such judgements.

An important area of practical application for the proposed prior ESS summaries is design and inference for early phase clinical trials with small to moderate sample sizes. A concern about inappropriately influential prior information is one of the main impediments to a more widespread use of Bayesian methods in clinical trials. With the ever growing pressure for efficient use of resources and higher ethical standards, clinical trial designs are becoming increasingly more complex. Trial designs now routinely include the use of multiple outcomes, borrowing strength across different disease subtypes or patient subpopulations, the use of biomarkers, adaptive allocation, sequential stopping, etc. Such complexity may lead naturally to the use of a CIHM. The ability to compute prior ESS in such models provides a natural basis for calibrating priors as well as explaining the model to non-statisticians. Prior ESS values provide a similar tool when using a CIHM in a meta-analysis, which is an area of extensive activity. The two case studies in Section 2 are typical examples. It may be argued that, without an understanding of the prior ESS, it is impossible to carry out a fair regulatory review of a clinical trial protocol using a method based on a Bayesian CIHM. The proposed methodology may be used to mitigate this concern.

Overall, we believe that reporting ESS is most useful in judging the relative role of the prior, either informative or non-informative, in a clinical trial design, and as a tool to greatly facilitate the process of design simulation to calibrate the prior and obtain OCs. In this setting, both investigators and regulators may be very concerned that the trial conduct should not be overly biased by (sometimes inappropriately) optimistic priors. The ESS also is helpful to report and interpret the results of a data analysis, since comparing the ESS to the data sample size allows one to judge the prior's relative informativeness.

Acknowledgments

Satoshi Morita's work was supported in part by Grant H20-CLINRES-G-009 from the Ministry of Health, Labour, and Welfare in Japan and a non-profit organization Epidemiological and Clinical Research Information Network (ECRIN). Peter Müller's work was partially supported by Grant NIH/NCI R01 CA 75981. Peter Thall's work was partially supported by Grant NIH/NCI R01 CA 083932. We thank the associate editor and the referees for their thoughtful and constructive comments and suggestions.

References

Berlin, J.A. and Colditz, G.A. 1999. The role of meta-analysis in the regulatory process for foods, drugs, and devices. *Journal of the American Medical Association* 281: 830-834.

Berry, D. A., and Stangl, D. K. 1996. *Bayesian Biostatistics.* New York: Marcel Dekker.

Congdon, P. 2005. *Bayesian Statistical Modelling* (*2nd Edition*). Chichester: John Wiley and Sons.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2004. *Bayesian Data Analysis* (*2nd Edition*). New York: Chapman and Hall/CRC.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1: 515-533.

Gray, R. J. 1994. A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics* 50: 244-253.

Kass, R. E. and Steffey, D. 1989. Approximate Bayesian Inference in Conditionally Independent Hierarchical Models. *Journal of the American Statistical Association* 84: 717-726.

Morita, S., Thall, P. F., Müller, P. 2008. Determining the effective sample size of a parametric prior. *Biometrics* 64: 595-602.

O'Quigley J, Pepe M, Fisher L. 1990. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 46: 33-48.

Spiegelhalter, D.J., Freedman, L.S. and Parmar, M.K.B. 1994. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Series A* 157: 357-416.

Stangl, D. K. 1996. Hierarchical Analysis of Continuous-Time Survival Models. In: Berry, D. A., and Stangl, D. K. eds. *Bayesian Biostatistics*: 429-450.

Thall, P. F., Wathen, K., 1, Bekele, B. N., Champlin, R. E., Baker, L. H., Benjamin, R. S. 2003. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes . *Statistics in Medicine* 22: 763-780.

Appendix

In Case 1 of a CIHM, the fact that constructing an ε-information prior $\pi_{1,0}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ for $\pi_1(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ ensures that $\pi_{12,0}(\boldsymbol{\theta} \mid \boldsymbol{\phi}) = \int \pi_{1,0}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\pi_2(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi})d\tilde{\boldsymbol{\theta}}$ is an ε-information prior for $\pi_{12}(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ can be proved as follows:

Assume for simplicity that $\bar{V}_j = \infty$; the proof in the case $\bar{V}_j < \infty$ is similar. Accounting for the variability at both levels of the CIHM, the variance of each θ_j can be written in the expanded form $\sigma^2_{\pi_{12}}(\theta_j) = E_{\pi_2}\{\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})\} + \sigma^2_{\pi_2}\{E_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})\}$. Similarly, using $\pi_{1,0}$ in place of π_1, $\sigma^2_{\pi_{12,0}}(\theta_j) = E_{\pi_2}\{\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\} + \sigma^2_{\pi_2}\{E_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\}$. Since the definition of ε-information prior ensures that $E_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0) = E_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})$, the second terms of the two expansions are identical. Denoting this common term by c_j, and also denoting $a_j = E_{\pi_2}\{\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})\}$ and $b_j = E_{\pi_2}\{\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\}$, the ratio of variances may be written as

$$\frac{\sigma^2_{\pi_{12}}(\theta_j)}{\sigma^2_{\pi_{12,0}}(\theta_j)} = \frac{E_{\pi_2}\{\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})\} + c_j}{E_{\pi_2}\{\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\} + c_j} = \frac{a_j + c_j}{b_j + c_j}. \tag{22}$$

Given $\tilde{\boldsymbol{\theta}}$, under the assumption $\bar{V}_j = \infty$ one can choose $\tilde{\boldsymbol{\theta}}_0$ so that $c_j/\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0) < \varepsilon/2$ and $\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})/\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0) < \varepsilon/2$. Writing these inequalities as $c_j - \sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\varepsilon/2 < 0$ and $\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}}) - \sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\varepsilon/2 < 0$, taking the expectations with respect to π_2 and reversing the algebra in each inequality, it follows that $c_j/E_{\pi_2}\{\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\} = c_j/b_j < \varepsilon/2$ and $E_{\pi_2}\{\sigma^2_{\pi_1}(\theta_j|\tilde{\boldsymbol{\theta}})\}/E_{\pi_2}\{\sigma^2_{\pi_{1,0}}(\theta_j|\tilde{\boldsymbol{\theta}}_0)\} = a_j/b_j < \varepsilon/2$. Writing the right-hand side of

(22) as $(a_j/b_j+c_j/b_j)/(1+c_j/b_j)$, since $c_j/b_j > 0$ this is bounded above by $a_j/b_j+c_j/b_j$, which by the foregoing is bounded above by $\varepsilon/2 + \varepsilon/2 = \varepsilon$.