STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2163–2182 (DOI: 10.1002/sim.855)

On determination of sample size in hierarchical binomial models

Kelly H. Zou[1,2,*,†] and Sharon-Lise T. Normand[1,3]

[1]*Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.*
[2]*Department of Radiology, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, U.S.A.*
[3]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.*

SUMMARY

We consider a two- and a three-stage hierarchical design containing the effects of k clusters with n units per cluster. In the two-stage model, the conditional distribution of the discrete response Y_i is assumed to be independent binomial with mean $n\theta_i$ $(i = 1, \ldots, k)$. The success probabilities, θ_i's, are assumed exchangeable across the k clusters, each arising from a beta distribution. In the three-stage model, the parameters in the beta distribution are assumed to have independent gamma distributions. The size of each cluster, n, is determined for functions of θ_i. Lengths of central posterior intervals are computed for various functions of the θ_i's using Markov chain Monte Carlo and Monte Carlo simulations. Several prior distributions are characterized and tables are provided for n with given k. Methods for sample size calculations under the two- and three-stage models are illustrated and compared for the design of a multi-institutional study to evaluate the appropriateness of discharge planning rates for a cohort of patients with congestive heart failure. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Over the past decade, multi-centre clinical trials have become increasingly utilized in the experimental setting. The National Institutes of Health in the United States, for example, have sponsored the formation of several multi-institutional collaborative groups to study treatments of cancer, AIDS and cardiovascular disease. In the non-experimental setting, similar trends in studies of the delivery and quality of medical care have been observed. Our work, in particular, is motivated by a study that involves the comparison of multiple health care providers on the basis of the quality of treatment rendered to patients having congestive heart failure (CHF). Complex studies such as these require the development of efficient study designs and data

*Correspondence to: Kelly H. Zou, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.
†E-mail: zou@hcp.med.harvard.edu

Contract/grant sponsor: Agency for Healthcare Research and Quality; contract/grant number: U18-HS09487

Received October 1999
Accepted September 2000

analyses. Although methodology for the analyses of multi-centre trials has grown substantially with the advent of inexpensive computing and convenient simulation algorithms that permit basic [1], multi-level [2, 3], empirical [4], hierarchical [5] and fully Bayes analyses [6], as well as stratification methods [7], methodology for the design of such trials is relatively limited.

In many multi-centre studies, in addition to the performance of any individual centre, interest is often focused on a particular aspect of the participating centres, such as the average treatment benefit or the range in treatment benefit. Thus the study objective can often be directed at estimation of a particular function of the centre-specific parameters across all centres. The goal of this article is to characterize the sample size necessary to make inference about a real-valued function of a rate parameter vector θ of length k, denoted by $\theta^* = t_k(\theta)$, over a range of constraints. Specifically, we consider the following hierarchical design containing the effects of k clusters and n cluster-specific units ($i = 1, \ldots, k$):

Stage I (*individual level*, *within-cluster model*). The conditional distribution of the discrete response Y_i is independent binomial with mean $n\theta_i$

$$(Y_i | n, \theta_i) \overset{\text{independent}}{\sim} \text{binomial}(n, \theta_i) \tag{1}$$

for $i = 1, \ldots, k$ clusters.

Stage II (*between clusters*). The success probability θ_i is assumed exchangeable across the k clusters, arising from a beta distribution with hyperparameters (α, β)

$$(\theta_i | \alpha, \beta) \overset{\text{iid}}{\sim} \text{beta}(\alpha, \beta) \tag{2}$$

for $i = 1, \ldots, k$, where $\alpha > 0$ and $\beta > 0$ are prespecified by the investigator. In the case of a single cluster, integration of θ from the joint distribution $f(Y_i, \theta_i \mid n, \alpha, \beta)$ leads to the beta-binomial (or Pólya) distribution.

A limitation of the model specified in equations (1) and (2) is the inability to learn about α and β by 'borrowing strength' from cluster to cluster, as in most hierarchical models. In order to learn about α and β, we also consider the addition of prior distributions for the parameters in the beta distribution:

Stage III (*common across all clusters*). The common hyperparameters α and β across all clusters have independent gamma distributions with parameters (p_α, q_α) and (p_β, q_β), respectively

$$(\alpha | p_\alpha, q_\alpha) \sim \text{gamma}(p_\alpha, q_\alpha) \perp (\beta | p_\beta, q_\beta) \sim \text{gamma}(p_\beta, q_\beta) \tag{3}$$

where all of the p's and q's take on positive values, with $p_\alpha \geqslant 1$ and $p_\beta \geqslant 1$ in order for the posterior to be log-concave. The main difference between the two- and three-stage models is that the third stage provides a hyperprior over the θ_i parameters that should provide for improved estimation when the θ's are distributed as assumed.

Sample size determination for a given study may be undertaken using a hypothesis-testing or an estimation approach, from either a frequentist or Bayesian framework. Frequentist approaches to designs involving a single rate parameter, θ, or a comparison of two rates, $\theta_1 - \theta_2$, are common. In the one-sample problem for θ, for example, sample size may be determined using the approximate mean and variance of the maximum likelihood estimate in a large-sample test (see Chapter 1.4 of Fleiss, reference [8], and references therein). In the case of

k clusters, Donner and colleagues [9, 10] derived and inverted a χ^2-test to obtain optimal cluster sample sizes when testing for differences among the k underlying rates. The authors developed solutions in the context of a two-group randomization design within each of the k clusters. In an unpublished manuscript by Chen, Weissfeld, and Ahnn of the University of Pittsburgh, the authors extended Donner's method to k-clusters for which units within a cluster are assigned to one of three treatment groups.

From the Bayesian perspective, several solutions have also been proposed. Joseph and colleagues [11, 12] proposed solutions for both the one-sample and two-sample problems. For a single θ, they considered the beta-binomial model specified in equations (1) and (2) and similar models were extended to compare two independent and dependent rates arising from possibly correlated beta prior distributions. In both cases, the researchers employed interval-based criteria using coverage of highest posterior density (HPD) regions. In contrast, Spiegelhalter and Freedman [13] and Spiegelhalter *et al.* [1] derived sample size for comparing two rates using the predictive power of a hypothesis test. They used the probability of a clinically important difference as the primary basis for their hypotheses. Hornberger and Eghtesady [14] developed methods for sample size calculation for a two-sample test by estimating the expected cost-benefit of a randomized trial. The prior probability of the success rates between the experimental and control groups were assumed to have a joint beta distribution, and sample size was determined by minimizing the posterior expected loss per patient in the study.

Methods for sample size determination in hierarchical designs are much less developed however. Parmigiani and Berry [15] proposed methods when interest centres on estimation of the hyperparameters in equation (2), assuming a three-stage model with the hyperparameters having distributions described in equation (3). The authors maximized the expected Lindley information [16] for the hyperparameters and derived the optimal choice of n and k. However, the method for determining sample size ultimately depends on the goal of the study; see, for example, the discussion in Joseph and Wolfson [17] regarding interval-based versus decision-analytic approaches.

In this article we extend the methods proposed by Joseph *et al.* [11, 12] to the hierarchical setting, developing methods to determine cluster size based on interval estimation of the composite parameter $\theta^* = t_k(\theta)$ for any given k. Because often at the planning phase no information other than prior event data are available to distinguish among the θ_i's, we assume exchangeability and apply interval-based criteria using coverage of central posterior intervals (CPI) to solve for optimal n. In Section 2, under the two-stage model, sample size determination is described using two estimation methods for the posterior distribution: direct simulation using Markov chain Monte Carlo methods and approximations based on sample moments. In Section 3, the direct simulation method is extended to the three-stage model. As recommended by Müller and Parmigiani [18], the sample size is determined by fitting a curve across the grid values of sample sizes for improved estimation. The authors developed numerical methods for the optimal design of a Bayesian Monte Carlo experiment. The authors proposed fitting a smooth utility surface and borrowing information from neighbourhood design points. We present the results of simulation studies describing the impact of the choice of prior distribution on sample size in Section 4. In Section 5 methods are illustrated on designing a multi-institutional study to determine the mean and range of appropriate hospital discharge planning rates for a cohort of CHF patients. We conclude with a discussion in Section 6.

2. A TWO-STAGE MODEL

Let the random variable $\mathbf{y}=\{y_1,\ldots,y_k\}$ represent the observed data from a sample space $\mathscr{Y}$, and similarly let $\theta=\{\theta_1,\ldots,\theta_k\}$ denote the vector of success probabilities assuming values in a parameter space Θ. Throughout we assume that the hyperparameters, (α,β), are prespecified by the investigator and the design is balanced with cluster sample sizes equal to n. Our general approach to sample size determination uses the 'average length criterion' suggested by Joseph *et al.* [12]. The minimum integer n such that the lengths of fixed coverage of CPIs averaged over the predictive distribution of the data is less than a prespecified value is found (see Appendix A). Formally, we find the minimum sample size n such that

$$\int_{l(\mathbf{y},n)}^{l(\mathbf{y},n)+\Delta'(\mathbf{y},n)} f(\theta^*\mid\mathbf{y},n)\,\mathrm{d}\theta^*=1-\delta \quad \text{with} \quad \int_{\mathscr{Y}}\Delta'(\mathbf{y},n)f(\mathbf{y})\,\mathrm{d}\mathbf{y}\leqslant\Delta \tag{4}$$

Here $f(\theta^*\mid\mathbf{y},n)$ is the posterior distribution for the composite parameter, $f(\mathbf{y})$ is the predictive distribution of $\mathbf{y}$, $l(\mathbf{y},n)$ is the lower limit of the CPI with length $\Delta'(\mathbf{y},n)$, and $1-\delta$ is the nominal coverage probability. The right-most term in equation (4) stipulates that the average of such intervals be no more than a fixed constant, Δ.

For the two-stage model specified by equations (1) and (2), the prior distributions are conjugate, and thus the posterior distribution of θ, given (α,β), is

$$f(\theta\mid\mathbf{y},\alpha,\beta)=c(\mathbf{y},n,\alpha,\beta)\prod_1^k\theta_i^{y_i+\alpha-1}(1-\theta_i)^{n-y_i+\beta-1} \tag{5}$$

where the constant $c(\mathbf{y},n,\alpha,\beta)=\prod_1^k\{\mathrm{B}(y_i+\alpha,n-y_i+\beta)\}^{-1}$ and $\mathrm{B}(\cdot,\cdot)$ is the beta function. The posterior distribution for θ^*, $f(\theta^*|\mathbf{y},\alpha,\beta)$ may be written as

$$\int_\theta|J(\theta\to\theta^*)|c(\mathbf{y},n,\alpha,\beta)\prod_1^k(t_{ki}^{-1})^{y_i+\alpha-1}(1-t_{ki}^{-1})^{n-y_i+\beta-1}\,\mathrm{d}f(\theta) \tag{6}$$

where J is the Jacobian of transformation from θ to θ^*, t_{ki}^{-1} is ith component of the inverse function t_k $(i=1,\ldots,k)$, and $f(\theta)$ is a function of the θ's to be integrated out, depending on the $t_k(\theta)$ considered. The predictive distribution is given by

$$f(\mathbf{y}|\alpha,\beta)=\prod_1^k\binom{n}{y_i}\mathrm{B}(y_i+\alpha,n-y_i+\beta)\{\mathrm{B}(\alpha,\beta)\}^{-1}$$

Because the posterior distribution in (6) is typically intractable, for any given n, numerical methods must be used to approximate the lengths of each CPI for $\Delta'(\mathbf{y},n)$ and then averaged over the predictive distribution. One wishes to find the minimum n such that the average length is within Δ.

For example, in the case of $k=2$ clusters with data $\mathbf{y}=(y_1,y_2)$ and unequal cluster sample sizes $\mathbf{n}=(n_1,n_2)$, suppose it is desired to construct an interval estimate of $\theta^*=t_k(\theta)=\theta_1-\theta_2$ within a specified length Δ, where

$$(\theta_i|\alpha_i,\beta_i)\overset{\text{independent}}{\sim}\text{beta}(\alpha_i,\beta_i)$$

for $i=1,2$. Because the difference is of interest, we assume prior parameters $\alpha=(\alpha_1,\alpha_2)$ and $\beta=(\beta_1,\beta_2)$ for the θ_i's. The posterior distribution becomes

$$f(\theta^*|\mathbf{y},\mathbf{n},\alpha,\beta)=\int_{\max(0,\theta^*)}^{\min(\theta^*+1,1)} f(\theta_1,\theta^*|\mathbf{y},\mathbf{n},\alpha,\beta)\,\mathrm{d}\theta_1 \tag{7}$$

where the integrand is

$$c(\mathbf{y},\mathbf{n},\alpha,\beta)\theta_1^{y_1+\alpha_1-1}(1-\theta_1)^{n_1-y_1+\beta_1-1}(\theta_1-\theta^*)^{y_2+\alpha_2-1}(1-\theta_1+\theta^*)^{n_2-y_2+\beta_2-1}$$

with $c(\mathbf{y},\mathbf{n},\alpha,\beta)=\prod_{i=1}^{2}\{\mathrm{B}(y_i+\alpha_i,n_i-y_i+\beta_i)\}^{-1}$. The quantity in (7) may be viewed as a polynomial in θ_1, and this integral can be solved analytically [12].

The investigator may also be interested in the average length of the CPIs for θ^*, over the predictive distribution, to have the fixed coverage probability of 95 per cent, such that the length is within Δ. The target sample sizes, n_1 and n_2, require solving

$$\int_{l(\mathbf{y},\mathbf{n})}^{l(\mathbf{y},\mathbf{n})+\Delta'(\mathbf{y},\mathbf{n})}\int_{\max(0,\theta^*)}^{\min(\theta^*+1,1)} f(\theta_1,\theta^*|\mathbf{y},\mathbf{n},\alpha,\beta)\,\mathrm{d}\theta_1\,\mathrm{d}\theta^*=0.95$$

with

$$\sum_{y_1}^{n_1}\sum_{y_2}^{n_2}\Delta'(\mathbf{y},\mathbf{n})\prod_{i=1}^{2}\binom{n_i}{y_i}\frac{\mathrm{B}(y_i+\alpha_i,n_i-y_i+\beta_i)}{\mathrm{B}(\alpha_i,\beta_i)}\leqslant\Delta$$

In the case of $k=3$ clusters with data $\mathbf{y}=(y_1,y_2,y_3)$ and an equal sample size n in each of these clusters, suppose it is desired to construct an interval estimate of $\theta^*=t_k(\theta)=(\theta_1+\theta_2+\theta_3)/3$ within a specified length Δ, where

$$(\theta_i|\alpha,\beta)\overset{\text{independent}}{\sim}\text{beta}(\alpha,\beta)$$

for $i=1,\ldots,3$. The posterior distribution becomes

$$f(\theta^*|\mathbf{y},n,\alpha,\beta)=\int_0^1\int_{\max(0,3\theta^*-\theta_2-1)}^{\min(3\theta^*-\theta_2,1)} f(\theta_1,\theta_2,\theta^*|\mathbf{y},n,\alpha,\beta)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2$$

where the integrand is

$$\begin{aligned}&c(\mathbf{y},n,\alpha,\beta)\theta_1^{y_1+\alpha-1}(1-\theta_1)^{n-y_1+\beta-1}\theta_2^{y_2+\alpha-1}(1-\theta_2)^{n-y_2+\beta-1}(3\theta^*-\theta_1-\theta_2)^{y_3+\alpha_3-1}\\&\quad\times(1-3\theta^*+\theta_1+\theta_2)^{n-y_3+\beta-1}\end{aligned}$$

with $c(\mathbf{y},n,\alpha,\beta)=\prod_{i=1}^{3}\{\mathrm{B}(y_i+\alpha,n-y_i+\beta)\}^{-1}$. The target sample size, n, requires solving

$$\int_{l(\mathbf{y},n)}^{l(\mathbf{y},n)+\Delta'(\mathbf{y},n)}\int_0^1\int_{\max(0,3\theta^*-\theta_2-1)}^{\min(3\theta^*-\theta_2,1)} f(\theta_1,\theta_2,\theta^*|\mathbf{y},n,\alpha,\beta)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\,\mathrm{d}\theta^*=0.95$$

with

$$\sum_{y_1}^{n}\sum_{y_2}^{n}\sum_{y_3}^{n}\Delta'(\mathbf{y},n)\prod_{i=1}^{3}\binom{n}{y_i}\frac{\mathrm{B}(y_i+\alpha,n-y_i+\beta)}{\mathrm{B}(\alpha,\beta)}\leqslant\Delta$$

We describe simulation methods to sample size determination in the context of a general number of clusters, k.

2.1. Sample size selection for general k using Markov chain Monte Carlo simulation

Direct simulations from the posterior distribution in (6) can be accomplished using Markov chain Monte Carlo methods. Because of conditional independence, Gibbs sampling, implemented in software such as BUGS [19], can be used to fit the models. In order to solve for n with fixed k, an initial guess n_0 is made by making an adjustment to the maximum likelihood estimate of n (see Appendix B).

The following algorithm for sample size calculation in the two-stage model is based on a grid search. The target CPI length and the nominal coverage probability $1-\delta$ are first specified. Around n_0, we select a grid of equal spacing. For various θ^* functions (for example, mean and range), the CPI lengths are computed by Monte Carlo simulations. Finally the optimal n satisfying the desired average CPI length within Δ is identified.

2.1.1. Sample size algorithm using Markov chain Monte Carlo sampling for θ^*. Assuming the number of clusters, k, is fixed, let n_0 denote the initial sample size, m the total number of Monte Carlo iterations, G the number of grid values around n_0, J the number of iterates of Gibbs' samplers, and $1-\delta$ the desired nominal coverage probability. To determine sample size:

1. Specify the hyperparameters (α, β), the number of clusters k, and the target average CPI length, Δ.
2. Calculate an initial guess for sample size, n_0. For example, if $\theta^* = \sum \frac{\theta_i}{k}$, then specify the initial guess by subtracting $(\alpha+\beta)$ from the maximum likelihood estimate of n

$$n_0 = \left\{ \frac{2\Phi^{-1}(\delta/2)}{\Delta(\alpha+\beta)} \right\}^2 \left(\frac{\alpha\beta}{k} \right) - (\alpha+\beta) \tag{8}$$

   where Φ is the cumulative distribution function of a standard normal distribution.
3. Create G grid values around n_0, yielding n_g for $g = 1, \ldots, G$ sample sizes.
4. For $l = 1, \ldots, m$ Monte Carlo iterations:
   (a) For $i = 1, \ldots, k$, generate $(\theta_{il}|\alpha,\beta) \sim \text{beta}(\alpha,\beta)$.
   (b) Calculate $(\theta_l^*|\alpha,\beta)$.
   (c) For $g = 1, \ldots, G$ grid values:
      (i) generate $(Y_{gil}|\theta_{il},\alpha,\beta) \sim \text{binominal}(n_{gl},\theta_{il})$;
      (ii) for $j = 1, \ldots, J$ iterates of Gibbs samplers, calculate $(\theta_{gl}^{*(j)}|\mathbf{y}_{gl},\alpha,\beta) = \frac{1}{k}\sum_{i=1}^{k}\{(\hat{\theta}_{gil}^{(j)}|gl,\alpha,\beta)\}$ by Markov chain Monte Carlo;
      (iii) calculate the CPI (I_{gl}) and length $\hat{\Delta}'(I_{gl})$ based on the $100(\delta/2)$ and $100(1-\delta/2)$ percentiles of the J ordered iterates $(\theta_{gl}^{*(j)}|\mathbf{y}_{gl},\alpha,\beta)$;
      (iv) compute coverage using the indicator $1\{(\theta_l^*|\alpha,\beta) \in I_{gl}\}$ where $1(\cdot)$ is the indicator function;
   (d) calculate Monte Carlo average length $\hat{\Delta}'_g(I) = \frac{1}{m}\sum_{l=1}^{m}\hat{\Delta}'(I_{gl})$;
   (e) calculate Monte Carlo coverage probability $\hat{\gamma}_g = \frac{1}{m}\sum_{l=1}^{m} 1\{(\theta_l^*|\alpha,\beta) \in I_{gl}\}$.

5. Repeat (b)–(d) for $g = 1, \ldots, G$ potential sample sizes. Choose $n'_g = \min_{n_g}\{\hat{\Delta}'_g(I) \leqslant \Delta\}$ as the required cluster size.

Note: in step 4(c)(ii), with independent beta priors, a more efficient approach is to directly simulate the J θ's from the posterior, followed by calculating θ^* for each of these.

2.2. Sample size selection using an approximation to the posterior distribution

Rather than simulating directly from the posterior distribution for the composite parameter, we approximate the posterior distribution in (5) by matching its moments to that of another distribution [20]. For simple composite functions of the parameters, such as the average, the advantage of this approach is that the CPIs are symmetric so that their lengths can be calculated explicitly. The ith marginal posterior distribution $(\theta_i|y_i, \alpha, \beta)$, conditioned on α and β, is

$$(\theta_i|y_i, \alpha, \beta) \overset{\text{independent}}{\sim} \text{beta}(c_i, d_i) \text{ with } c_i = y_i + \alpha \quad \text{and} \quad d_i = n - y_i + \beta \tag{9}$$

The posterior distribution $f(\bar{\theta}|\mathbf{y}, \alpha, \beta)$ can be approximated by a Normal distribution $\text{N}(m, v)$, where

$$m = E(\bar{\theta}|\mathbf{y}, \alpha, \beta) = \frac{1}{k}\sum_{i=1}^{k}\frac{c_i}{c_i + d_i}$$

and

$$v = \text{var}(\bar{\theta}|\mathbf{y}, \alpha, \beta) = \frac{1}{k^2}\sum_{i=1}^{k}\frac{c_i d_i}{(c_i + d_i)^2(c_i + d_i + 1)} \tag{10}$$

with c_i and d_i given in (9). Alternatively, because of skewness, the posterior distribution in (6) may also be approximated by a $\text{beta}(s, t)$ distribution, with

$$s = m\left\{\frac{m(1-m)}{v} - 1\right\} \quad \text{and} \quad t = (1-m)\left\{\frac{m(1-m)}{v} - 1\right\} \tag{11}$$

and (m, v) given in (10).

2.2.1. Sample size determination using an approximation to the posterior distribution for θ^*. The algorithm for sample size determination is similar to that described in Section 2.1.1. Step 4(c)(ii) of the original algorithm is replaced by

(c)(ii) Calculate CPI $I_{gl} = m_l \pm \Phi^{-1}(\delta/2)(v_l)^{1/2}$ and length $\hat{\Delta}'(I_{gl}) = 2\Phi^{-1}(\delta/2)(v_l)^{1/2}$, where (m_l, v_l) are given in (10)

in the case of a Normal approximation to the posterior, or by

(c)(ii) Calculate CPI I_{gl} and length $\hat{\Delta}'(I_{gl})$ based on a sorted random sample generated by $\text{beta}(s_l, t_l)$, given in (11)

in the case of a beta approximation.

2.3. Specifying the prior distribution

In order to design a study, a prior distribution for the rate parameters must be specified by the investigator, which subsequently impacts sample size determination. The choices of priors can

be based on several criteria (see the discussion by Kass and Wasserman [21] on guidelines for the selection of prior distributions). Subjective prior distributions can be created from pilot data or could be elicited from experts. A moment-approach or matching a given functional form [22] is the most frequently used method. The investigator can choose the prior density that most closely matches prior beliefs.

In the two-stage model, the prior mean and variance of $(\theta_i|\alpha,\beta)\sim\text{beta}(\alpha,\beta)$ are used to determine α and β ($\alpha>0$; $\beta>0$). Symmetry ($\alpha=\beta$) or lack of it ($\alpha\neq\beta$) can be suitably characterized. A uniform prior is $(\alpha,\beta)=(1,1)$ [23].

Although most Bayesian analyses are performed with non-informative priors constructed by formal rules, such as Jeffreys' prior [24] with $(\alpha,\beta)=(0.5,0.5)$, it is not necessary to adopt such a prior when dealing with simple hierarchical binomial models. This is especially true at the design stage, where very different sample sizes are needed for estimating binomial parameters near 0 or 1, compared to values near 0.5.

3. A THREE-STAGE MODEL

In order to learn about α and β, we consider the additional third stage stated in equation (3). The two independent gamma distributions of the third-stage hyperprior distributions are not conjugate with the first two. The joint posterior distribution of θ, given $(p_\alpha, p_\beta, q_\alpha, q_\beta)$, is

$$f(\theta|\mathbf{y}, p_\alpha, q_\alpha, p_\beta, q_\beta)=\int_0^\infty\int_0^\infty f(\alpha|p_\alpha,q_\alpha)f(\beta|p_\beta,q_\beta)f(\theta|\mathbf{y},\alpha,\beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta \tag{12}$$

where $f(\alpha|p_\alpha,q_\alpha)$ and $f(\beta|p_\beta,q_\beta)$ have two independent gamma distributions, and $f(\theta|\mathbf{y},\alpha,\beta)$ is given by (5). Similarly, the predictive distribution can be obtained.

More so than the two-stage model, the posterior distribution of a function of the θ_i's is quite intractable. In addition, the approximation methods are not possible because of the non-conjugacy of the third-stage hyperprior distribution. Therefore, we employ direct sampling from the posterior distribution by Markov chain Monte Carlo. The algorithm is similar to that presented in Section 2.1, except that $(p_\alpha,q_\alpha,p_\beta,q_\beta)$ are specified. For an initial guess of n_0, we approximate the second-stage parameters by $(\alpha_0|p_\alpha,q_\alpha)=p_\alpha/q_\alpha$ and $(\beta_0|p_\beta,q_\beta)=p_\beta/q_\beta$. Consequently, n_0 is obtained by (8), where α and β are replaced by $(\alpha_0|p_\alpha,q_\alpha)$ and $(\beta_0|p_\beta,q_\beta)$, respectively.

In order to reduce the computational burden from fitting the three-stage model, while still providing accurate sample sizes, we adopt the smoothing procedure of Müller and Parmigiani [18] by smoothing the observed CPI lengths against the corresponding evenly spaced grid-values of sample sizes. In practice, we employ the 'lowess' smoother in S-plus with smoothing parameter f.

3.1. Prior distributions in the three-stage model

Because the variance of a gamma(p,q) distribution is p/q^2, the inverse-scale hyperparameters, q_α and q_β, determine the precision of the prior information. In the absence of information about (α,β), we seek a relatively diffuse hyperprior distribution by specifying small q_α and q_β. Conversely, large values for the q's provide more informative priors. The expectation of a gamma(p,q) distribution is p/q, allowing us to derive *ad hoc* shape hyperparameters,

p_α and p_β. For example, if a uniform beta$(1,1)$ distribution is used in a two-stage model, we make $p_\alpha = q_\alpha$ and $p_\beta = q_\beta$ in a three-stage model, so that the underlying expectations of these gamma distributions are both 1. However, we recommend updating the priors using pilot data, whenever available, after fitting a three-stage model to pilot data using *ad hoc* hyperparameters selected above. See the example in Section 5 for details on eliciting and updating the hyperprior distributions.

4. SIMULATION STUDIES

We investigated the average lengths of CPIs for several composite parameters $t_k(\theta)$ using Monte Carlo simulation. Our simulations studies were based on a full-factorial design involving four factors: (1) number of clusters (four levels), $k = \{10, 20, 50, 100\}$; (2) sample size per cluster (five levels), $n = \{5, 20, 50, 100, 500\}$; (3) type of composite parameter (five levels), $\theta^* = \{\text{mean}, \text{median}, \text{min}, \text{max}, \text{range}\}$; and (4) type of prior distribution (5 levels) in the two-stage model, $(\alpha, \beta) = \{(1, 3), (1, 1), (9, 9), (3, 1), (9, 1)\}$, so that $E(\theta|\alpha, \beta)$ were $\{0.25, 0.50, 0.50, 0.75, 0.90\}$. In order to have approximately comparable means of θ_i to the two-stage model, we used the following parameters in the three-stage model: $(p_\alpha, q_\alpha, p_\beta, q_\beta) = \{(9, 9, 9, 3), (9, 9, 9, 9), (9, 1, 9, 1), (9, 3, 9, 9), (9, 1, 9, 9)\}$.

The nominal coverage probability, $1 - \delta$, was set at 95 per cent. Fifty ($m = 50$) Monte Carlo iterations were used to estimate all CPI lengths. We applied both direct simulation and approximation methods for estimating the posterior distribution for the composite parameter.

For each Monte Carlo iteration and under each of the five prior distributions in the two-stage models, we first generated 100 rate parameters, $(\theta_i|\alpha, \beta)$. We then generated the corresponding data $\mathbf{y}(n, \theta)$ for various n's. The first $k \in \{10, 20, 50, 100\}$ of the 100 $(\mathbf{y}|\theta, \alpha, \beta)$'s and the true $(\theta|\alpha, \beta)$ were used in the study. We used direct simulation based on 500 iterates after a burn-in of 500 iterates, employing the Gibbs sampler as well as the method of matching moments. The latter method also utilized 50 simulations to estimate the posterior distribution. Similar simulation procedures were conducted under each of the corresponding five priors in the three-stage model.

Tables of CPI lengths were constructed with fixed k, while cluster sizes, n, were permitted to vary. Although we studied sample size for the mean, median, minimum, maximum, and range of the rate parameters, we report results only for the mean and range (the remaining results are available from the authors).

4.1. Results

Selected results using the direct simulation method are reported in Tables I and II for the mean and range functions, respectively. For completeness, Table III presents the large-sample 95 per cent frequentist confidence intervals (see Appendix B). As expected, the average CPI lengths in Table I were smaller than those reported in Table II.

For both the mean and range functions, the three-stage models generally yielded greater average CPI lengths than the two-stage models, especially when $n < 50$. Such differences also depended on the value of the third-stage hyperpriors. For example, $(p_\alpha, q_\alpha, p_\beta, q_\beta) = (9, 1, 9, 1)$ gave much larger variances in their gamma distributions, compared with $(9, 9, 9, 9)$. As a

Table I. Estimated lengths of 95 per cent CPI intervals ($\hat{\Delta}$) for the mean function, $\frac{1}{k}\sum_i \theta_i$, in the two- and three-stage models.

Two-stage models: (α, β)																				
	$E(\theta_i)=0.25$				$E(\theta_i)=0.50$				$E(\theta_i)=0.50$				$E(\theta_i)=0.75$				$E(\theta_i)=0.90$			
	(1, 3)				(1, 1)				(9, 9)				(3, 1)				(9, 1)			
	k				k				k				k				k			
n	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100
5	0.162	0.112	0.071	0.050	0.189	0.136	0.086	0.061	0.123	0.092	0.059	0.038	0.158	0.112	0.071	0.050	0.092	0.065	0.041	0.029
20	0.099	0.070	0.044	0.031	0.106	0.077	0.048	0.035	0.096	0.067	0.044	0.029	0.098	0.069	0.044	0.031	0.064	0.046	0.029	0.021
50	0.065	0.046	0.029	0.021	0.069	0.050	0.032	0.022	0.072	0.052	0.033	0.022	0.065	0.046	0.029	0.021	0.045	0.032	0.021	0.014
100	0.047	0.033	0.021	0.015	0.049	0.036	0.023	0.016	0.054	0.039	0.026	0.014	0.046	0.033	0.021	0.015	0.033	0.024	0.015	0.010
500	0.021	0.015	0.009	0.006	0.022	0.015	0.010	0.007	0.026	0.018	0.012	0.008	0.021	0.016	0.009	0.007	0.015	0.010	0.007	0.005
Three-stage models: $(p_\alpha, q_\alpha, p_\beta, q_\beta)$																				
	$E(\theta_i)=0.25$				$E(\theta_i)=0.50$				$E(\theta_i)=0.50$				$E(\theta_i)=0.75$				$E(\theta_i)=0.90$			
	(9, 9, 9, 3)				(9, 9, 9, 9)				(9, 1, 9, 1)				(9, 3, 9, 9)				(9, 1, 9, 9)			
	k				k				k				k				k			
n	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100
5	0.190	0.140	0.091	0.064	0.204	0.145	0.095	0.068	0.230	0.169	0.115	0.080	0.183	0.143	0.094	0.067	0.121	0.098	0.064	0.047
20	0.099	0.073	0.047	0.033	0.105	0.077	0.048	0.034	0.123	0.088	0.058	0.040	0.096	0.076	0.048	0.034	0.072	0.053	0.034	0.024
50	0.063	0.047	0.030	0.021	0.067	0.048	0.031	0.022	0.081	0.058	0.037	0.026	0.063	0.048	0.031	0.022	0.047	0.034	0.022	0.015
100	0.045	0.033	0.021	0.015	0.047	0.034	0.022	0.015	0.058	0.041	0.026	0.018	0.045	0.034	0.022	0.016	0.033	0.024	0.016	0.011
500	0.020	0.014	0.009	0.007	0.021	0.015	0.010	0.007	0.026	0.018	0.012	0.008	0.021	0.015	0.010	0.007	0.015	0.011	0.007	0.005

Table II. Estimated lengths of 95 per cent CPI intervals ($\hat{\Delta}$) for the range function, range $\{\theta_i\}$, in the two- and three-stage models.

Two-stage models: (α, β)

	$E(\theta_i)=0.25$				$E(\theta_i)=0.50$				$E(\theta_i)=0.50$				$E(\theta_i)=0.75$				$E(\theta_i)=0.90$			
	$(1,3)$				$(1,1)$				$(9,9)$				$(3,1)$				$(9,1)$			
	k				k				k				k				k			
n	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100
5	0.455	0.391	0.303	0.248	0.340	0.211	0.099	0.050	0.320	0.284	0.235	0.208	0.453	0.394	0.313	0.254	0.320	0.316	0.292	0.277
20	0.317	0.285	0.237	0.202	0.234	0.164	0.085	0.047	0.289	0.257	0.221	0.196	0.321	0.288	0.240	0.210	0.257	0.254	0.247	0.234
50	0.228	0.208	0.179	0.155	0.169	0.124	0.069	0.037	0.238	0.218	0.193	0.175	0.233	0.210	0.182	0.158	0.193	0.194	0.196	0.188
100	0.174	0.153	0.138	0.121	0.120	0.094	0.055	0.032	0.195	0.156	0.137	0.132	0.173	0.161	0.137	0.124	0.156	0.153	0.153	0.150
500	0.085	0.076	0.069	0.062	0.057	0.047	0.030	0.021	0.104	0.098	0.092	0.083	0.083	0.076	0.066	0.061	0.082	0.079	0.078	0.070

Three-stage models: $(p_\alpha, q_\alpha, p_\beta, q_\beta)$

	$E(\theta_i)=0.25$				$E(\theta_i)=0.50$				$E(\theta_i)=0.50$				$E(\theta_i)=0.75$				$E(\theta_i)=0.90$			
	$(9,9,9,3)$				$(9,9,9,9)$				$(9,1,9,1)$				$(9,3,9,9)$				$(9,1,9,9)$			
	k				k				k				k				k			
n	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100
5	0.473	0.404	0.322	0.262	0.386	0.250	0.143	0.084	0.351	0.329	0.292	0.266	0.473	0.385	0.311	0.248	0.356	0.348	0.332	0.310
20	0.322	0.290	0.231	0.199	0.253	0.174	0.108	0.068	0.302	0.278	0.251	0.226	0.315	0.276	0.224	0.190	0.266	0.262	0.250	0.235
50	0.223	0.202	0.171	0.149	0.177	0.124	0.084	0.054	0.244	0.224	0.201	0.183	0.224	0.200	0.162	0.139	0.200	0.194	0.196	0.186
100	0.173	0.153	0.132	0.116	0.129	0.095	0.064	0.044	0.197	0.178	0.164	0.151	0.170	0.149	0.126	0.110	0.153	0.152	0.151	0.145
500	0.083	0.075	0.066	0.059	0.063	0.047	0.033	0.025	0.102	0.097	0.090	0.086	0.082	0.072	0.062	0.057	0.079	0.074	0.074	0.075

Table III. Large-sample frequentist 95 per cent confidence intervals ($\hat{\Delta}$) for the mean function, $\frac{1}{k}\sum_i \theta_i$.

	$E(\theta_i)=0.50$				$E(\theta_i)=0.75$ (or 0.25)				$E(\theta_i)=0.90$			
	k				k				k			
n	10	20	50	100	10	20	50	100	10	20	50	100
5	0.277	0.196	0.124	0.088	0.240	0.170	0.107	0.076	0.166	0.118	0.074	0.053
20	0.139	0.098	0.062	0.044	0.120	0.085	0.054	0.038	0.083	0.059	0.037	0.026
50	0.088	0.062	0.039	0.028	0.076	0.054	0.034	0.024	0.053	0.037	0.024	0.017
100	0.062	0.044	0.028	0.020	0.054	0.038	0.024	0.017	0.037	0.026	0.017	0.012
500	0.028	0.020	0.012	0.009	0.024	0.017	0.011	0.008	0.017	0.012	0.007	0.005

result, greater differences of the average lengths were observed between $(p_\alpha, q_\alpha, p_\beta, q_\beta)=(9,\ 1,\ 9,\ 1)$ in the three-stage model and $(\alpha,\beta)=(9,9)$ in the two-stage model, than between $(9,\ 9,\ 9,\ 9)$ and $(\alpha,\beta)=(9,\ 9)$.

When estimation of the mean rate across clusters is of interest, the direct simulation and approximation methods yielded almost identical results. However, the approximation methods required much less computing time and may be preferred. In all cases we examined, the beta approximation did not have any extra advantage over the Normal approximation. In general, the extremely skewed priors yielded the smallest average length. Furthermore, the 'prior sample size' of the beta prior distribution (that is, the sum of the two prior beta parameters) plays an important role in determining this length.

When estimation of the range in cluster-specific rates is of interest, the two slightly skewed priors generally yielded greatest and almost identical average lengths for $k=10$. The uniform prior had the smallest average lengths for a large number of clusters ($k=100$). Figure 1 displays the average 95 per cent CPI lengths for the range of the rates with 10 (upper) and 100 (lower) clusters, and with two- (left) and three-stage (right) prior distributions.

Finally, all methods are accurate with coverage results approximately equal to the nominal coverage level.

5. EXAMPLE: CONGESTIVE HEART FAILURE

The Q-SPAN Cardiovascular Disease study, funded by the Agency for Healthcare Research and Quality, involves the examination of quality of care rendered to patients with cardiovascular conditions, one of which is congestive heart failure (CHF). An important indicator of the quality of care subsequent to hospitalization is whether a patient has an appropriate discharge plan. The investigators wanted to design a study to estimate the average rate of appropriate hospital discharge planning and the range in these rates across a consortium of $k=30$ hospitals.

5.1. A two-stage design

In the absence of prior studies, the investigators assumed that the mean rate was 0.75 and that the average length of a CPI for the range of rates was 3.5 times greater than that for the

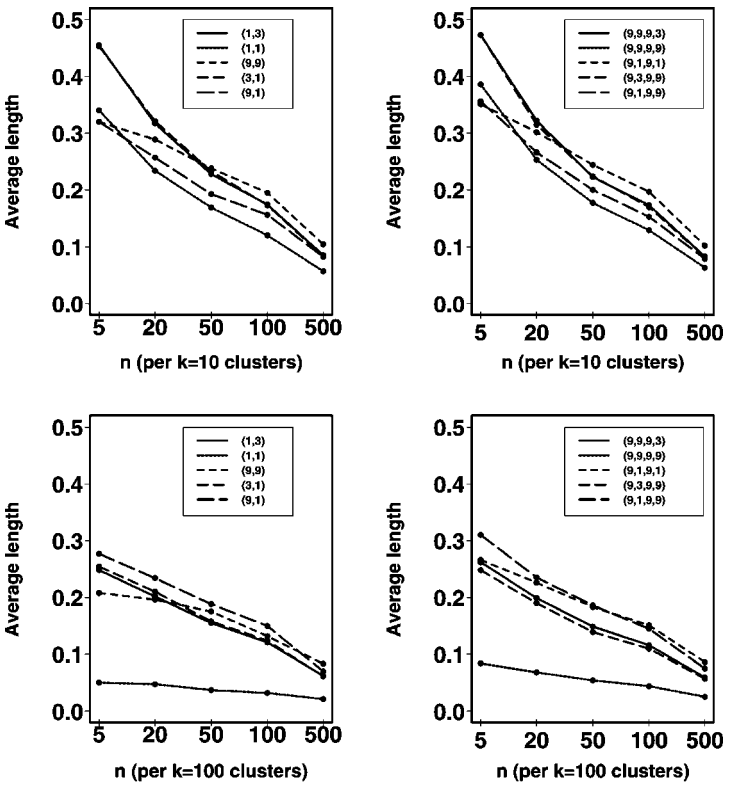

Figure 1. Average 95 per cent CPI length for the range of rates for $k = 10$ (upper) and 100 (lower) and for the two- (left) and three-stage (right) binomial models.

mean rate. Using these assumptions, a beta prior with hyperparameters $(\alpha, \beta) = (29, 10)$ (see the Appendix C) was constructed.

Figures 2 and 3, respectively, display the required hospital sample sizes, n, when interest centres on the mean rate and range in rates. Using a predetermined length of 2.5 per cent for the average, the required sample sizes are 109, 110 and 114 by the direct simulation, Normal approximation, and beta approximation methods for the composite parameter. In contrast, 154 patients per hospital are required if estimation is to proceed by the frequentist large-sample method. Given a CPI length of at most 10 per cent for the range in discharge planning rates, 230 patients per hospital are required.

5.2. *A three-stage design*

The investigators obtained pilot information from 21 hospitals, each hospital having a minimum of five patients meeting entry criteria. For each patient, information abstracted from medical records regarding the appropriateness of the discharge plan was available. To determine sample size for the new study, a three-stage model with unequal cluster sizes, n_i, was fitted to the pilot data assuming $(p_\alpha, q_\alpha, p_\beta, q_\beta) = (54, 1, 7, 1)$. The posterior means of the rates

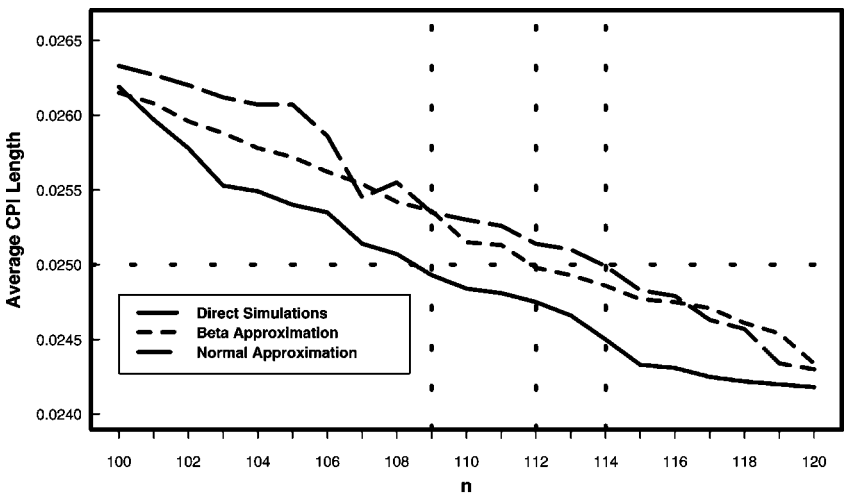

Figure 2. Average 95 per cent CPI length for the mean rate for $k = 30$ in a two-stage model, using the prior beta(29, 10).

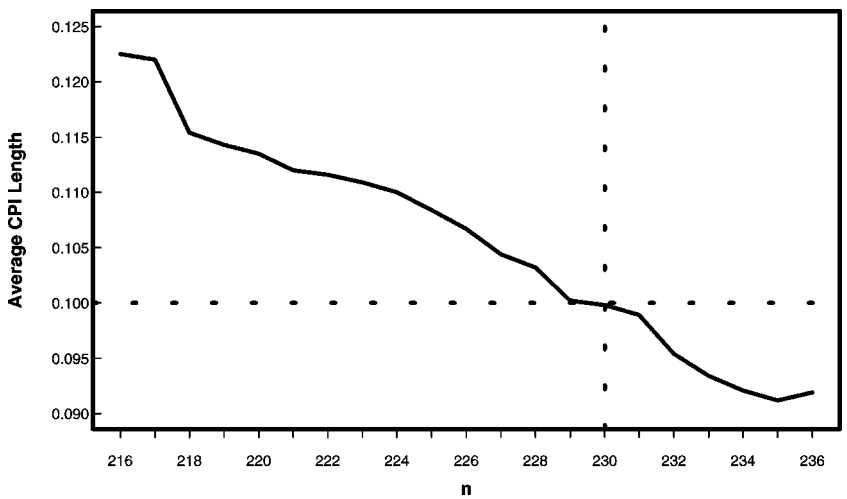

Figure 3. Average 95 per cent CPI length for the range of the rates for $k = 30$ in a two-stage model, using the prior beta(29, 10).

of the 21 hospitals ranged between 0.87 to 0.91. Based on these results, the hyperparameters were updated (see Appendix D), resulting in $(p_\alpha, q_\alpha) = (57, 1)$ and $(p_\beta, q_\beta) = (16, 3)$, which were used to determine the number of CHF patients per hospital necessary for the new study of 30 hospitals.

Figures 4 and 5 display the required hospital sample sizes, n. A grid using cluster sizes ranging between 20 to 65 in increments of 5 was created. A lowess smoother was applied to

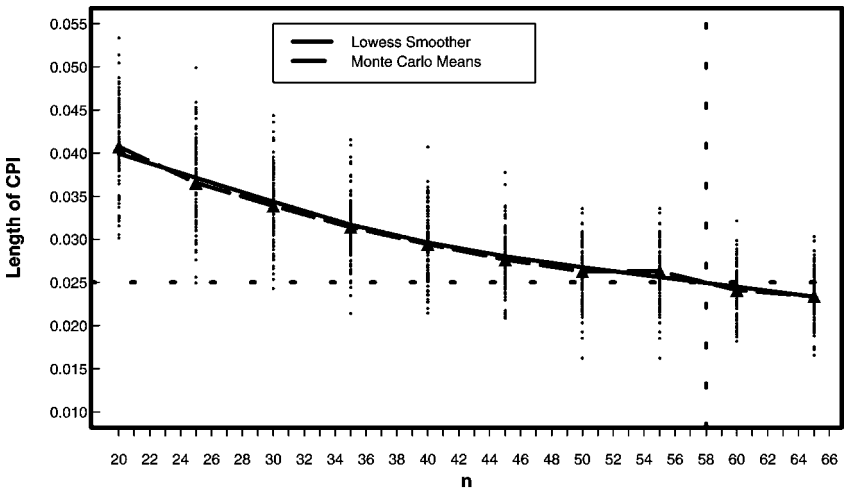

Figure 4. Average 95 per cent CPI length for the mean rate for $k = 30$ in a three-stage model, using two independent priors gamma$(57, 1)$ and gamma$(16, 3)$.

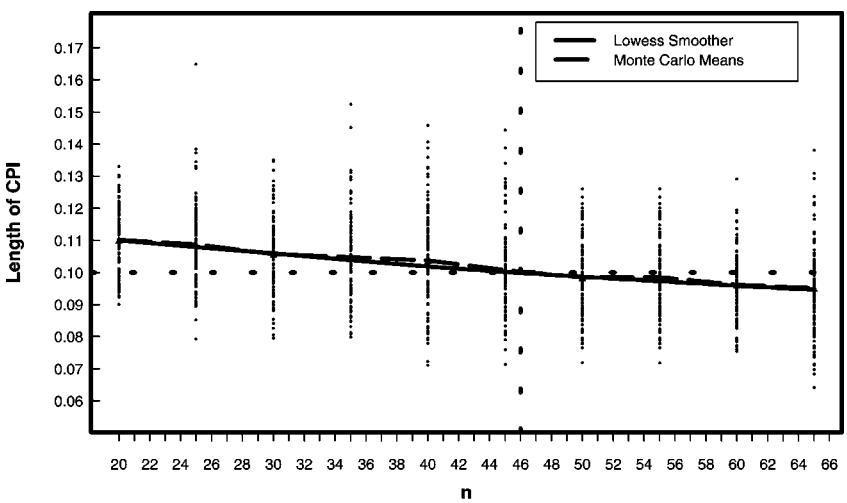

Figure 5. Average 95 per cent CPI length for the range of the rates for $k = 30$ in a three-stage model, using two independent priors gamma$(57, 1)$ and gamma$(16, 3)$.

the scatter plot of CPI lengths at these grid values with the degree of smoothing, f, ranging from 0.30 to 0.90. The results suggest that given a CPI length of at most 2.5 per cent for the mean rate, 58 patients per hospital are required while 46 patients per hospital are required to achieve a CPI length of at most 10 per cent for the range in discharge planning rates. The required sample sizes were robust to choice of f.

6. DISCUSSION

Practical methods to implement designs in hierarchical models are limited although the use of such designs has increased in both experimental and observational settings. In this article we considered hierarchical binomial models with two or three stages. We examined various functions of the rate parameters across the clusters. Algorithms for computing the number of patients per cluster, n, for given number of clusters k, over a range of priors, were developed and illustrated with an example.

Both two- and three-stage models produced smaller average CPI lengths than those without assuming any of the hierarchical structures. Compared with the two-stage models, the three-stage models yielded greater average CPI lengths for the same approximated true rates, especially for small n in our specific simulations. It may not always be the case that the three-stage model will provide equal or greater sample sizes relative to the two-stage model as the sample size depends on the choice of prior distributions. The advantage of the three-stage model is also to allow for borrowing strength from neighbour design clusters and for updating the prior distributions whenever pilot information is available. We demonstrated in our example that much smaller sample sizes were needed by considering a three-stage model and pilot information.

The choice of the prior distribution impacts the average CPI length. In the two-stage models, when the mean function was of interest, all priors gave similar average lengths, especially when cluster sizes were large ($n \geqslant 100$). For estimation of the range of the rates, the uniform prior gave greater average lengths when the number of clusters was small. The opposite was true when the number of clusters was large. This may be because the lower and upper bounds of the CPIs tend to be close when we sample with these two priors. In the three-stage model, the precision of the third-stage hyperprior distributions impacts the differences in the average CPI length between the two- and three-stage models.

It is not surprising that the beta and Normal approximation methods in the two-stage design are easy to implement and performed very well. The approximation using a beta distribution did not add extra accuracy in estimation over the range of parameters examined. In more complex situations, Markov chain Monte Carlo may be needed.

We approached the sample size problem from an interval-based framework. Some authors have argued that this approach has the drawback of not including costs [17]. With the increasing concern of the cost-effectiveness issue when designing a health services research study, it may be advantageous to revisit the Bayesian decision and information approach in which a utility function can be specified and maximized for the functions of the parameters of interest.

It is also important to note that different approaches to sample size determination can lead to different answers, even in the context of an interval-based approach. Joseph *et al.* [11] determined samples sizes when requiring the average interval lengths to be less than a cut-off while fixing the coverage probability ('average length criterion') and compared to sample sizes when requiring the average coverage probability of fixed interval lengths to maintain a prespecified level ('average coverage criterion'). They suggested using the former because it typically requires fewer subjects, especially when the nominal coverage level is high or when the desired average CPI length is small. Although we did not perform an extensive simulation study based on the average coverage criterion, we did perform some simulations and observed results similar to those reported by Joseph and colleagues when using a two-stage design.

Several extensions are necessary to this research on several fronts. Two important extensions involve use of other sampling models, such as the Poisson or Normal and inclusion of covariates in both the sampling model and in the prior distribution [25, 26]. The latter extension is especially needed in the observational setting where case-mix will vary across the clusters.

APPENDIX A: BAYESIAN POSTERIOR COVERAGE PROBABILITY

Let $I_\delta(\mathbf{y}) = \{l(\mathbf{y}, n), l(\mathbf{y}, n) + \Delta(\mathbf{y}, n)\}$ be an interval estimate of θ^* based on $\mathbf{y}$ for fixed values of the hyperparameters α and β. The performance of this interval is measured by averaging the posterior probability that θ^* is included in $I_\delta(\mathbf{y})$ over the distribution of the data. We denote this quantity by $Q(\mathbf{y}) = \int_{\theta^*} 1\{\theta^* \in I_\delta(\mathbf{y})\} \times f(\theta^* \mid \mathbf{y})\, \mathrm{d}\theta^*$, where 1(.) is an indicator function assuming a value of 1 if the condition is true and 0 otherwise. A $100(1-\delta)$ per cent CPI for θ^* is defined as a subset $I_\delta(\mathbf{y})$ of all $\theta^* \in \Theta$ such that $1 - \delta \leqslant Q(\mathbf{y})$. In repeated draws of $\{\mathbf{y}, \theta^*\}$ from $f(\mathbf{y} \mid \theta^*) f(\theta^*)$, the average posterior probability coverage of $I_\delta(\mathbf{y})$ is thus given by (Rubin and Schenker [27])

$$\gamma\{\theta^*\} = \int_{\mathbf{y} \in \mathscr{Y}} \int_{\Theta} Q(\mathbf{y}) f(\mathbf{y} \mid \theta^*) f(\theta^*)\, \mathrm{d}\theta^*\, \mathrm{d}\mathbf{y} \tag{A1}$$

APPENDIX B: SAMPLE SIZE DETERMINATION BY THE MAXIMUM LIKELIHOOD METHOD

Ignoring the second stage in the model, assume a common underlying $\theta_i = \theta$, and a balanced design of size n for each cluster. By large-sample approximation, for the average function, $t_k(\theta) = \bar{\theta} = \theta_i$, the desired length of $100(1-\delta)$ per cent confidence interval for $\bar{\theta}$ is

$$\Delta = 2\Phi^{-1}(\delta/2) \left\{ \frac{\theta(1-\theta)}{kn} \right\}^{1/2}$$

where Φ^{-1} is the standard normal quantile. The resulting n for any given Δ is then

$$n = \frac{4\left\{\Phi^{-1}(\delta/2)\right\}^2 \theta(1-\theta)}{k[\Delta\{I_\delta(\bar{\theta})\}]^2} \tag{B1}$$

Note that this solution does not accommodate between-cluster heterogeneity. In practice, however, θ in (14) is substituted by $E(\theta|\alpha, \beta) = \alpha/(\alpha + \beta)$ if there is a second-stage beta prior with hyperparameters of (α, β).

APPENDIX C: METHOD OF MOMENTS FOR DETERMINING THE BETA HYPERPARAMETERS IN THE EXAMPLE IN SECTION 5.1

Let (m, v) be the underlying mean and variance of the second-stage parameter θ_i $(i = 1, \ldots, k)$, where

$$(\theta_i|\alpha, \beta) \overset{\text{iid}}{\sim} \text{beta}(\alpha, \beta)$$

so that $m = \alpha/(\alpha+\beta)$ and $v = \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. Denote the 'standardized mean' as $c_1 = m/v$. It can easily be obtained that

$$\alpha = m(\alpha+\beta) \quad \text{and} \quad \beta = \left(\frac{1-m}{m}\right)\alpha$$

For k clusters, the underlying mean and variance of the cluster average $\bar{\theta}$, respectively, are

$$E(\bar{\theta}|\alpha,\beta) = E(\theta_i|\alpha,\beta) = m$$

and

$$\text{var}(\bar{\theta}|\alpha,\beta) = \frac{1}{k^2}\sum_{i=1,k}\text{var}(\theta_i|\alpha,\beta) = \frac{v}{k} = \frac{m}{c_1 k}$$

We assume that the posterior of the mean rate parameter over k clusters is an estimator with above underlying mean and variance. Using the frequentist large-sample central limit theorem as a crude approximation, a 95 per cent 'confidence interval' of $(\bar{\theta}|a,b)$ has length

$$I_{0.05}(\bar{\theta}|a,b) = 2 \times 1.96 \times \left(\frac{m}{c_1 k}\right)^{1/2}$$

leading to

$$c_1 = \frac{3.92^2 m}{k\{I_{0.05}(\bar{\theta}|\alpha,\beta)\}^2} \tag{C1}$$

In addition, we assume that the ratio of the length of 'confidence intervals' of the range and sample average of the estimated rate parameters over clusters is

$$c_2 = I_{0.05}\{\text{range}(\theta)|\alpha,\beta)\}/I_{0.05}(\bar{\theta}|\alpha,\beta) \tag{C2}$$

for $c_2 \geqslant 1$. Thus, from (C1) and (C2),

$$c_1 = \frac{(3.92)^2 m}{k\{I_{0.05}(\bar{\theta}|\alpha,\beta)\}^2} = \frac{(3.92)^2 m c_2^2}{k[I_{0.05}\{\text{range}(\theta)|\alpha,\beta\}]^2} \tag{C3}$$

If the range is of interest, we only need to specify m, k, $I\{\text{range}(\theta)|\alpha,\beta\}$ and c_2. Alternatively, if the mean of interest we need to specify the m, k and $I_{0.05}(\bar{\theta}|\alpha,\beta)$. We then may use (C3) for the initial (α,β)-value. This procedure can be repeated by trial and error for the desirable c_2-value.

In our case, we assume $m = 0.75$, $k = 30$, $c_2 = 3.5$, and $I_{0.05}\{\text{range}(\theta)|\alpha,\beta\} = 0.10$, translating to $I_{0.05}(\bar{\theta}|\alpha,\beta) = 0.03$, approximately. Thus by (C3)

$$c_1 = \frac{3.92^2 \times 0.75 \times 3.5^2}{30 \times 0.1^2} = 471$$

resulting in

$$\alpha = 0.75 \times (471 \times 0.25 - 1) = 29 \quad \text{and} \quad \beta = \frac{1-0.75}{0.75} \times 29 = 10$$

Note that in the final results we give the sample size n needed for the target length of CPI credible set $I_{0.05}(\bar{\theta}|\mathbf{y},\alpha,\beta) = 0.025$, close to the above 0.03 assumption.

APPENDIX D: METHOD OF MOMENTS FOR DETERMINING THE GAMMA HYPERPARAMETERS IN THE EXAMPLE IN SECTION 5.2

Let (m_α, v_α) be the underlying mean and variance of the third-stage parameter α, where

$$(\alpha | p_\alpha, q_\alpha) \sim \text{gamma}(p_\alpha, q_\alpha)$$

where $m_\alpha = p_\alpha/q_\alpha$ and $v_\alpha = p_\alpha/q_\alpha^2$. It follows that

$$p_\alpha = m_\alpha^2/v_\alpha \quad \text{and} \quad q_\alpha = m_\alpha/v_\alpha \tag{D1}$$

Estimate (m_α, v_α) by their respective posterior mean and variance from the pilot data, and then update the hyperparameters (p_α, q_α) by equation (D1). Similarly we apply this method to derive (p_β, q_β).

In our case, we first fit the three-stage model using the pilot data and hyperparameters $(p_\alpha, q_\alpha, p_\beta, q_\beta) = (54, 1, 7, 1)$. The resulting posterior mean and variance for α were 53.91 and 51.24. Thus

$$p_\alpha = 53.91^2/51.24 = 57 \quad \text{and} \quad q_\alpha = 53.91/51.24 = 1$$

The posterior mean and variances for β were 6.44 and 2.52. Thus

$$p_\beta = 6.44^2/2.52 = 16 \quad \text{and} \quad q_\beta = 6.44/2.52 = 3$$

ACKNOWLEDGEMENTS

This work was supported by grant U18-HS09487 from the Agency for Healthcare Research and Quality, Rockville, MD. Dr Zou completed this research while a Postdoctoral Fellow in Statistics in the Department of Health Care Policy, Harvard Medical School. We acknowledge constructive comments from Dr Barbara J. McNeil of Harvard Medical School. We also thank two anonymous referees for their careful reviews of earlier versions.

REFERENCES

1. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *Journal of Royal Statistical Society*, *Series A* 1994; **157**:357–416.
2. Goldberg JD, Koury KJ. Design and analysis of multicenter trials. In *Statistical Methodology in the Pharmaceutical Sciences*, Berry DA (ed). Marcel Dekker: New York, 1990; 201–237.
3. Goldstein H. *Multilevel Statistical Models*. Halstead Press: New York, 1995.
4. Morris CN. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 1983; **78**:47–65.
5. Berry DA. A Bayesian approach to multicenter trials and meta-analysis. *ASA Proceedings of the Biopharmaceutical Section* 1990: 1–10.
6. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Chapman & Hall: London, 1995; 128–134.
7. Brandmaier RM, Aydemir Ü, Ansari H, Hasford J. Do we always need clinical stratification in multicenter clinical trials? Results of a simulation study. *Proceedings of the SAS Users Group International Conference* 1992; **17**:1439–1444.
8. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley: New York, 1981; 13–17.
9. Donner A, Birkett N, Buck C. Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology* 1981; **114**:906–914.
10. Donner A. Sample size requirement for stratified cluster randomization designs. *Statistics in Medicine* 1985; **11**:743–750.
11. Joseph L, Wolfson DB, du Berger R. Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician* 1995; **44**:143–154.

12. Joseph L, du Berger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* 1997; **16**:769–781.
13. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* 1986; **5**:1–13.
14. Hornberger J, Eghtesady P. The cost-benefit of a randomized trial to a health care organization. *Controlled Clinical Trials* 1998; **19**:198–211.
15. Parmigiani G, Berry DA. Applications of Lindley information measure to the design of clinical experiments. In *Aspects of Uncertainty*, Freeman PR, Smith AFM (eds). Wiley: New York, 1994.
16. Lindley DV. Binomial sampling schemes and the concept of information. *Biometrika* 1957; **44**:179–186.
17. Joseph L, Wolfson DB. Interval-based versus decision theoretic criteria for the choice of sample size. *Statistician* 1997; **46**:145–149.
18. Müller P, Parmigiani G. Optimal design via curve fitting of Monte Carlo experiment. *Journal of the American Statistical Association* 1995; **90**:1322–1330.
19. Spiegelhalter D, Thomas A, Best N, Gilks, W. *BUGS*: *Bayesian inference using Gibbs sampling*, *version 0.5*. MRC Biostatistics Unit: Cambridge, 1996.
20. Tanner MA. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Function*, 2nd edn. Springer-Verlag: New York, 1993; 9–19.
21. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 1996; **91**:1343–1370.
22. Berger JO. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag: New York, 1985; 78–81.
23. Villegas C. On the representation of ignorance. *Journal of the American Statistical Association* 1977; **72**:651–654.
24. Jeffreys H. *Theory of Probability*, 3rd edn. Oxford University Press: London, 1961.
25. Joseph L, Belisle P. Bayesian sample size determination for normal means and differences between normal means. *Statistician* 1997; **46**:209–226.
26. Snijders TA, Bosker RJ. Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 1993; **18**:237–259.
27. Rubin DB, Schenker N. Efficiently simulating the coverage properties of interval estimates. *Applied Statistics* 1986; **35**:159–167.

