

This article was downloaded by:[Kristofferson, Eric]  
On: 19 April 2008  
Access Details: [subscription number 792304965]  
Publisher: Taylor & Francis  
Informa Ltd Registered in England and Wales Registered Number: 1072954  
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Journal of Biopharmaceutical Statistics

Publication details, including instructions for authors and subscription information:  
<http://www.informaworld.com/smpp/title~content=t713597263>

### Experience with Reviewing Bayesian Medical Device Trials

Gene Pennello <sup>a</sup>; Laura Thompson <sup>a</sup>

<sup>a</sup> Division of Biostatistics, Rockville, Maryland, USA

Online Publication Date: 01 January 2008

To cite this Article: Pennello, Gene and Thompson, Laura (2008) 'Experience with Reviewing Bayesian Medical Device Trials', Journal of Biopharmaceutical Statistics, 18:1, 81 - 115

To link to this article: DOI: 10.1080/10543400701668274

URL: <http://dx.doi.org/10.1080/10543400701668274>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## EXPERIENCE WITH REVIEWING BAYESIAN MEDICAL DEVICE TRIALS

Gene Pennello and Laura Thompson

Division of Biostatistics, Rockville, Maryland, USA

*The purpose of this paper is to present a statistical reviewer's perspective on some technical aspects of reviewing Bayesian medical device trials submitted to the Food and Drug Administration. The discussion reflects the experiences of the authors and should not be misconstrued as official guidance by the FDA. A variety of applications are described, reflecting our experience with therapeutic and diagnostic devices. In addition to Bayesian analysis of trials, Bayesian trial design and Bayesian monitoring are discussed. Analyses were implemented in WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>), with the code provided.*

**Key Words:** Center for Devices and Radiological Health (CDRH); Hierarchical model; Historical control; Interim analysis; Markov Chain Monte Carlo (MCMC); Model checking; Sample size determination; Sensitivity analysis; Study design; Subgroup analysis.

### 1. INTRODUCTION

Bayesian medical device trials have become an integral part of a significant proportion of premarket submissions to the Food and Drug Administration (FDA). The Bayesian initiative began about eight years ago within the Center for Devices and Radiological Health (CDRH), led by Division of Biostatistics Director Gregory Campbell and then Center Director Bruce Burlington. At CDRH, companies are encouraged to take advantage of good prior information on the safety and effectiveness of their investigational devices through the formal use of Bayesian analysis.

Over the years, CDRH has accumulated a body of institutional knowledge on important lessons in the design, conduct, and analysis of Bayesian medical device studies under regulatory review. Some of this thinking is shared in the document "Draft FDA Guidance on the Use of Bayesian Statistics for Medical Devices Trials" (<http://www.fda.gov/cdrh/osb/guidance/1601.pdf>), released in May 2006 for public comment. The guidance is written for multiple audiences, e.g., scientists and regulatory affairs personnel as well as statisticians.

The purpose of this paper is to present a statistical reviewer's perspective on some technical aspects of reviewing Bayesian medical device trials. The discussion reflects the experiences of the authors and should not be misconstrued as official

Received March 5, 2007; Accepted June 20, 2007

Address correspondence to Laura Thompson, Division of Biostatistics, 1350 Piccard Drive HFZ-550, Rockville, MD 20850, USA; E-mail: [laura.thompson@fda.hhs.gov](mailto:laura.thompson@fda.hhs.gov)

guidance by the FDA. A variety of applications are described, reflecting our experience with therapeutic and diagnostic devices. The Bayesian models used were implemented in the commonly used software package WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>). The code is provided in Appendix I. Hierarchical models are used extensively to combine prior information with study data and for structuring related parameters.

In the remainder of Section 1, general remarks are given on opportunities for Bayesian analysis of medical device trials and on the use of prior information. In subsequent sections, more technical discussions ensue on assessment of Bayesian trial design (Section 2), Bayesian sample size determination (Section 3), leveraging historical controls (Section 4), subgroup analysis (Section 5), Bayesian monitoring (Section 6), model checking (Section 7), and Bayesian methods for diagnostic devices (Section 8). The paper concludes with brief remarks on additional topics (Section 9).

### 1.1. Why Bayesian Medical Device Trials?

Device trials can be particularly suitable for Bayesian analysis. For example, if a therapeutic device has evolved in relatively small increments from previous generations of the same type of device, then prior information from the trials of the previous devices can be predictive of the safety and effectiveness profile of the new device. The reason the previous trials can be predictive is that the mechanism of action of a therapeutic device is often physical, implying a local effect that is often predictable. In contrast, the mechanism of action of pharmaceuticals is pharmacokinetic/pharmacodynamic, implying systemic effects that are often unpredictable from similar but not identical formulations. Other potentially reliable sources of prior information for device trials include clinical trials of the device conducted overseas, patient registries, pilot studies, studies of the device on similar patient populations, and perhaps nonclinical studies. Historical controls can also represent prior information for the control arm of a randomized controlled trial (see Section 4).

Besides enabling the formal use of prior information, the Bayesian approach can also be a flexible alternative to the frequentist (classical) approach to monitoring a trial (see Section 6). For example, stopping the trial early for unplanned reasons can create difficulties in interpreting or even defining frequentist  $p$ -values because they depend on the sample space. Even prospectively designed adaptive trials can cause controversy over the  $p$ -value, as exemplified by a trial of the extracorporeal membrane oxygenation (ECMO) treatment (Begg, 1990). In principle, Bayesian inferences are still valid in these circumstances because they depend instead on the parameter space. Moreover, the Bayesian approach is flexible at handling interim analyses and modifications to trials, such as dropping an unfavorable treatment arm or a modification to the randomization scheme. However, to assess the frequency with which one would make an incorrect decision for an adaptively conducted Bayesian trial, the frequentist operating characteristics (Type I error rate, power) can be very useful.

Other uses of Bayesian analysis include the ability to do complex modeling tasks, and to obtain a valid analysis when standard frequentist analyses do not exist or rely on suspect asymptotic approximation. For example, Bayesian analysis can be advantageous for small sample sizes, rare events, or when the maximum likelihood

estimate is on the boundary of the parameter space. A complex modeling task for which a Bayesian analysis might be advantageous is incorporation of multiple sources of uncertainty via random effects, mixed effects, or mixture modeling. Other tasks include Bayesian multiplicity adjustment (Sections 5, 6, and 9), adjustment for missing data, or sensitivity analysis for informative missingness or unmeasured covariates (Section 7.2). Bayesian methods have also been used to handle difficult analysis issues with diagnostic devices (Section 8).

## 1.2. The Use of Prior Information

Our experience is that the evaluation of prior information and its usefulness in a proposed study is an important scientific exercise that may involve clinicians, medical officers, engineers, and other scientists from the company, the FDA, and academia. Such experts can assess whether the prior information is “valid” in the sense that it is thought to provide clinically relevant information for the proposed study.

Prior information for a proposed study is usually based on previously conducted studies on the same device or a similar device. Generally, for the previous studies to be considered appropriate as prior information, they need to be similar to the proposed study in a number of factors, which can include objective, device used, clinical endpoints studied, covariates measured, study protocol, patient population, patient management, time and place the studies were conducted, physician training and experience, and technique used to deploy the device (Irony and Pennello, 2001). Previous studies that are proposed as prior information should also be representative of the possible studies that could have been selected to avoid selection bias.

An entirely subjective approach toward formulation of the prior distribution may be problematic. For example, consider a premarket application (PMA) for a device that is to be presented to an Advisory Panel of outside scientific experts. Such panels assist FDA on decisions of approval or disapproval. Some or all of the Advisory Panel members may strongly disagree with the selection of the subjective prior. In contrast, a quantitative approach to formulating the prior centers the debate on the quality of the studies used as prior information and not on post hoc reliance of subjective prior information after the data from the current study have been revealed.

It is important that the ramifications of the choice of prior information and its modeling be clearly appreciated. Device companies need to understand that if the prior data turn out to be sufficiently distant from the current study data, so that in retrospect the prior data were poor predictors for the current study, a Bayesian approach could actually be disadvantageous in that the weight of the prior information will have to be overcome by new data. The degree to which the prior data could be disadvantageous can be quantified through simulation and can help a company determine if a Bayesian approach is in its best interest.

Bayesian statistical approaches are sufficiently different from frequentist ones. Many experienced statisticians may not be well trained in the foundations and the computational aspects of a Bayesian design and analysis, Bayesian methodology, and the use of associated statistical software. A successful company generally has a

solid Bayesian statistician (or someone who really wants to learn) as an employee or consultant.

## **2. ASSESSMENT OF BAYESIAN TRIAL DESIGN: INFLUENCE OF PRIOR INFORMATION**

Like frequentist trials, a Bayesian device trial is designed to be able to demonstrate a hypothesis, or claim, with regard to a primary safety or effectiveness endpoint. To assess the design of a Bayesian trial, we have found special metrics to be insightful for assessing the influence of the prior information on the analysis (Sections 2.1 and 2.2). Besides these special metrics, calculation of Type I error rate and power are also frequently helpful in understanding trial design (Section 2.3).

### **2.1. Prior Probability of a Claim**

When combining prior information with the current study using a Bayesian model, a basic measure of the influence of the prior information is the prior probability of the claim on the primary endpoint. If the prior probability of the claim is too large, then the prior information is ordinarily discounted via some modification to the Bayesian model.

For example, consider a hierarchical model for combining data on a parameter in a study with prior information on the same parameter from previous studies. At the first level, the hierarchical model allows the parameter value to vary from study to study, but assumes that the study-specific parameter values are exchangeable, i.e., they are identically distributed, random draws from a common distribution, called a hyper-distribution. At the second level, the model assumes that observations within each study are exchangeable, i.e., (again) random draws from a common distribution. The parameters are related in that information on one provides information on the common hyper-distribution, which in turn provides information on the other parameters. Consequently, when computing the (prior or posterior) probability of the claim on the parameter in the current study, the hierarchical model “borrows strength” from information in the previous studies.

Hierarchical models are flexible in that the less consistent the prior information is with the current study data, the less it is used to estimate the current study parameter. However, when several previous studies are combined with the current study, the previous studies can attain a certain momentum, such that even if they are inconsistent with the current study, they can still have large influence on the posterior distribution of the current study parameter. This momentum can manifest itself as a large prior probability of the claim.

For a hierarchical model, as with other complex Bayesian models, the prior probability of the claim is not readily apparent. However, Gibbs sampling or another posterior sampling method can compute it. To illustrate, consider a hypothetical one-armed, 200-patient study of a cardiovascular device. The primary endpoint is 30-day MACE rate, a composite of major adverse cardiac events. We remark that although randomized controlled studies are generally preferred, one-armed studies can be acceptable for some classes of devices. Suppose the primary claim is that the 30-day MACE rate for the device is less than 0.249, a clinically

accepted objective performance criterion, and it has been agreed that the study is considered to be successful at demonstrating the claim if its posterior probability is greater than 0.975.

Thirty-day MACE rate data on six previous generations of the device are proposed as prior information (Table 1). After review, the six previous generations and the new device are considered similar from an engineering standpoint, and the six studies of the previous generations are considered similar to the one-armed study in endpoint, study protocol, patient population, and time frame. Therefore, the six previous studies are taken to be prior information and all seven studies are assumed exchangeable on 30-day MACE rate. We remark that adjustment for patient-level covariates is ordinarily needed for exchangeability of studies to be plausible because the studies are usually not identical in covariate distributions. However, for the sake of simplicity we ignore this in the modeling.

A binomial-normal hierarchical model can be used to model the number of subjects  $y_s$  in study  $s$  suffering a MACE event within 30 days after placement of the device among  $n_s$  total subjects, and the underlying study-specific 30-day MACE rates  $p_s$ ,  $s = 1, \dots, 7$ , that are assumed to be exchangeable:

$$\begin{aligned} y_s &\sim \text{Bin}(n_s, p_s) \\ \text{logit}(p_s) = \mu_s &\sim N(\mu_0, \sigma_\mu^2), \\ \mu_0 &\sim N(0, 1000), \\ \sigma_\mu^{-2} &\sim \Gamma(0.001, 0.001) \end{aligned} \tag{1}$$

Diffuse hyperpriors are placed on the mean  $\mu_0$  and precision  $\sigma_\mu^{-2}$  of the exchangeable logit parameters so that these are essentially estimated from the data alone. Let  $s = 1$  index the new study and  $s = 2, 3, \dots, 7$  index the previous studies. The parameter of interest is  $p_1$ , the 30-day MACE rate for the new study. The claim to be shown is  $p_1 < 0.249$ . The prior probability of the claim is defined as that probability conditional on  $y_2 - y_7$  but not on  $y_1$ , the data on the new device yet to be observed. In the software package WinBUGS, the prior distribution for  $p_1$  and therefore the prior probability of the claim are easily computed via Gibbs sampling by simply sampling a new proportion parameter in logit form from the logit hyper-distribution.

**Table 1** Hypothetical data on the 30-day MACE rate for six historical studies (HS) and one new study of a cardiovascular device

Device	$N$	Events	Rate
New	200	NA	NA
HS1	135	20	0.15
HS2	260	55	0.21
HS3	1960	325	0.17
HS4	415	60	0.15
HS5	205	43	0.21
HS6	25	5	0.20

**Table 2** Summary of prior distribution for 30-day MACE rate for the new device

Parameter	Mean	SD	2.5%	97.5%
$p_1$	0.176	0.034	0.121	0.252
$0.249 - p_1$	0.073	0.034	-0.003	0.128
$p_1 < 0.249$	0.973	0.163	0.0	1.0

Under the hierarchical model, the prior probability of the claim  $p_1 < 0.249$  is 97.3% (Table 2). The high prior probability results because the six previous studies are fairly consistent in the 30-day MACE rate. This consistency leads to a small estimate of the between-study variance  $\sigma_\mu^2$  in the logit of the true MACE rate, which in turn leads the model to borrow substantially from the prior information when estimating  $p_1$ , even before data on  $p_1$  have been observed.

The high prior probability of the claim suggests that the assumption of exchangeability of the studies is too strong to be warranted. The prior probability of the claim nearly exceeds the posterior probability success criterion of 0.975. That is, under the hierarchical model, effectiveness of the device has almost been demonstrated, without the need for the 200-patient clinical study. This conclusion would be at odds with a predetermination that a clinical study of the device is required, i.e., prior information alone is not enough to support approval of the device.

The hierarchical model can be modified to downweight the prior information such that the prior probability of the claim is lowered to an acceptable level. An ad hoc modification is to change the hyperprior on the between study precision  $\sigma_\mu^{-2}$  from  $\sigma_\mu^{-2} \sim \Gamma(0.001, .001)$  to  $\sigma_\mu^{-2} \sim \Gamma(10, 10)$ . As a result, the prior probability of the claim decreases from 0.973 to 0.674 (Table 3), a more reasonable level. Under the initial diffuse hyperprior, the between-study precision is estimated from the previous studies to be 220.7 (the posterior mean conditional on  $y_2 - y_7$  only). By comparison, the alternative hyperprior has a much smaller mean precision 1 with small variance 0.1. This hyperprior forces the model to yield a much smaller estimate of the between study precision than 220.7, leading to the desired result of less borrowing from the prior information when estimating  $p_1$ .

The level at which the prior probability of the claim should be set relies on clinical and engineering expertise. In this hypothetical example, the level 0.674 could have been elicited based on evaluation of the similarity of the previous generation devices with the new device on engineering characteristics and their

**Table 3** Summary of prior distribution for 30-day MACE rate for the new device after discounting the prior information

Parameter	Mean	SD	2.5%	97.5%
$p_1$	0.215	0.155	0.028	0.608
$0.249 - p_1$	0.034	0.155	-0.359	0.221
$p_1 < 0.249$	0.674	0.469	0.0	1.0

presumed impact on 30-day MACE rate. As mentioned in section 1, this subjectivity could be a source of controversy, especially if the device is taken to an FDA Advisory Panel. However, all analysis plans contain some degree of subjectivity. For example, the definition of study success (e.g., 30-day MACE rate target level 0.249) is a prespecification that is scientifically based but at its roots is a subjective determination.

## 2.2. Effective Sample Size

The effective sample size is a special metric for quantifying how many extra patients the prior distribution was worth. The effective sample size for a Bayesian analysis of a parameter of interest is the sample size at which the analysis is effectively operating when it combines prior information with study data on that parameter. Specifically, the effective sample size for parameter  $\theta$  is

$$ESS = N \frac{\text{Var}(\theta | \text{data, prior information ignored})}{\text{Var}(\theta | \text{data, prior information utilized})} \quad (2)$$

where  $N$  is the actual sample size of the study and the denominator and numerator are the posterior variances of  $\theta$  when prior distributions are used that utilize and ignore the prior information, respectively (Malec, 2001). The formula is motivated by the notion that, roughly speaking, variance is inversely proportional to sample size.

In the cardiovascular device example (Table 1), if 40 MACE events were observed among the 200 patients, then, under the binomial-normal hierarchical model expressed in Equation (1), the posterior mean for the MACE rate  $p_1$  in the current study is 0.184 and the posterior standard deviation is 0.018686. However, if instead the diffuse prior distribution  $\text{logit}(p_1) = \mu_1 \sim N(0, 1000)$  was used, then the posterior mean is 0.200 and the posterior standard deviation is 0.02826. The effective sample size is therefore  $200 \cdot (0.02826/0.018686)^2 = 457.4$ . That is, in addition to the sample size of 200 patients for the study, another 257.4 patients were effectively borrowed from the pool of 3000 patients in the six studies making up the prior distribution. Because these are 57.4 more patients than in the study itself, the prior information may be thought to be too informative. That concern was already indicated by a high prior probability of the claim  $p_1 < 0.249$  (Table 2).

As can be seen from the example, the effective sample size is a useful metric that is easy to understand. It is often communicated to clinicians and engineers on review teams to help facilitate their understanding of the ramifications of a Bayesian analysis.

At the design stage of a study, a preposterior analysis of the effective sample size can be made. An average effective sample size can be computed by averaging  $ESS$  over the prior distribution and the data. The computation can be made by simulation. The simulation algorithm is to generate parameters according to the prior distribution, generate data given the parameters, compute the effective sample size based on posterior analysis, and repeat the process to obtain a distribution for effective sample size over which the average is taken.

We remark that the effective sample size as defined here should not be confused with the effective sample size as defined in the Markov chain Monte Carlo

literature. There, effective sample size refers to the number of samples generated in a Gibbs sampling or other Markov chain posterior sampling method after accounting for auto-correlation in the samples.

### 2.3. Type I Error and Power

Although frequentist properties of Bayesian designs are not routinely evaluated for academic research, our experience is that FDA can request that Type I error rate and power be assessed for a proposed trial design. After all, the mandate of the FDA with regard to premarket approval applications is to rely on valid scientific evidence indicating reasonable assurance of safety and effectiveness. Therefore, it is not unreasonable that the regulatory agency would be concerned with the frequency with which it makes correct decisions.

Customarily, the assessment of frequentist type I error rate involves simulating the trial many times at the boundary of the null hypothesis, keeping track of the number of times the null hypothesis is rejected. In symbols, given null hypothesis  $H$ , for parameter vector  $\underline{\theta} \in H$  the type I error rate is  $\alpha(\underline{\theta}) = \int_R f(\underline{x} | \underline{\theta}) d\underline{x}$ , where  $f(\underline{x} | \underline{\theta})$  is the density for the data  $\underline{x}$  and  $R$  is the rejection region of the decision rule, which could be Bayesian or frequentist.

In the cardiovascular example (Table 1), calculation the Type I error rate is not too difficult. The boundary of the null hypothesis is  $p_1 = 0.249$ . Under binomial-normal hierarchical model, the study success criterion ( $p_1 < 0.249$  with posterior probability at least 0.975) is achieved if the number of MACE events observed is as high as 44 among the 200 patients in the study. Exploiting that the posterior probability is monotonically decreasing in the event count  $y_1$ , the Type I error rate is  $\Pr(y_1 \leq 44 | n = 200, p_1 = 0.249)$ , the cumulative distribution of  $y_1$  evaluated at 44, where  $y_1$  has binomial distribution  $Bin(200, 0.249)$ . Therefore, the one-sided Type I error rate is 0.194. Under the alternative hyperprior  $\Gamma(10, 10)$  for between study precision  $\sigma_\mu^{-2}$ , the success criterion is achieved with 38 or fewer events, leading to a one-sided Type I error rate of 0.0297. Power can be computed similarly for assumed values of  $p_1 < 0.249$  (see Section 3.3).

As an extension, a preposterior analysis can be made of the average type I error rate  $\alpha^* = \int \alpha(\underline{\theta}) \pi(\underline{\theta} | H) d\underline{\theta}$ , which integrates the type I error rate  $\alpha(\underline{\theta})$  over  $\underline{\theta}$  according to  $\pi(\underline{\theta} | H) = I(\underline{\theta} \in H) \pi(\underline{\theta}) / \int_H \pi(\underline{\theta}) d\underline{\theta}$ , the prior density conditional on  $H$  being true. The average type I error rate can be relevant if the prior information is highly reliable and likely to be repeated in the current trial. However, our experience is that in its mission to protect the public health, FDA, in its evaluation of Bayesian study designs, can make allowance for the possibility of a proposed trial being worse than expected, even if the prior information is considered highly reliable beforehand. Therefore, in addition to or in lieu of the average type I error rate, the usual type I error rate can be computed at specific parameter values in  $H$ , especially those considered clinically egregious.

When comparing frequentist and Bayesian designs, it is not unreasonable to maintain the same standard for level of evidence. For Bayesian designs, if the prior information is objective, highly reliable, and agreed upon by all parties, then the level of evidence can be regarded as being based on the study data and the prior information taken together. The implication is that a Bayesian design can maintain the same standard for level of evidence as a frequentist trial, yet have a type I

error rate larger than the nominal level of 5%. To see this in a simplified context, consider a case where the prior information is a  $z$  statistic that is large, indicating strong support in favor of the device. Then, in order to reject the null hypothesis the  $z$  statistic for the study data need not have to be as large as it would have to be if it was considered alone. The  $z$  statistic for the study data only needs to be large enough such that when combined with the  $z$  statistic for the prior information (in accordance with an agreed-upon Bayesian model), the combined test statistic meets the same evidence standard as would be required in a frequentist trial (e.g.,  $z$  statistic  $> 1.96$ ). Essentially, we can still maintain a type I error rate at a nominal level of say 5%, when it is calculated by considering both new and prior data together. Appendix II elaborates, showing for a particular example that when prior information is regarded as data and is taken together with the study data, the 5%-level frequentist decision rule is identical to the Bayesian decision rule based on a 95% posterior probability threshold.

### 3. SAMPLE SIZE DETERMINATION

The Bayesian approach can be viewed as sequential in nature, in which the posterior distribution is continually updated with new data until a conclusion is reached. However, the logistics in conducting a clinical study usually require some notion of sample size, regardless of whether the analysis will be Bayesian or frequentist. Sequential sample size or interim analysis designs will be discussed in Section 6. In this section, we focus on fixed sample size study designs, which have been used because they can be easier to implement for some device trials.

Bayesian approaches to sample size determination can be merely inferential or can involve optimal decision analysis with respect to a loss function. An inferential approach could be based on satisfying a posterior performance criterion or on the Bayes factor for choosing between models. Wang and Gelfand (2002) review the literature on performance criterion and Bayes factor approaches for sample size determination, illustrating the simulation steps necessary for determining the sample size. Pezeshk (2003) reviews all three approaches (decision analysis, posterior performance, and Bayes factor) from a clinical trials perspective.

#### 3.1. Performance Criteria

Performance criteria on which sample size has been based include preposterior analysis of average posterior variance, average coverage for a posterior interval (e.g., central or highest posterior density interval) with fixed length, average length for a posterior interval with fixed coverage, and average posterior probability that a parameter is greater than a target value (Wang and Gelfand, 2002). A power approach to Bayesian sample size determination has been considered for coronary stent device trials by O'Malley et al. (2002). Here, we also focus on power as the determinant for sample size.

For clinical studies under regulatory review, our experience is that the claim to be shown is often that a parameter is greater than a target value. For such a claim, the decision rule is usually that the claim is shown if its posterior probability exceeds  $1 - \gamma$ , for some specified  $\gamma$  (e.g., 0.025). For sample size determination,

the performance criterion that the average posterior probability is required to be  $1 - \gamma$  is in our experience usually not enough, because it implies a power to show the claim of only about 50%. (It is exactly 50% if the median posterior probability is  $1 - \gamma$ .) A sufficient sample size would require higher power, say, 80%.

### 3.2. Algorithm for Sample Size Determination

Bayesian sample size is predicated on two types of priors, a design prior reflecting one's prior belief about parameter values and a fitting prior that is to be used for analysis. The two types of priors are used in a fashion that is analogous to the frequentist approach to sample size determination. In the frequentist approach, specific values for parameters are assumed (a degenerate design prior) when computing the power to detect the alternative hypothesis using a prespecified analysis. However, the analysis does not assume anything about the parameter values (akin to a Bayesian analysis with a diffuse fitting prior).

The Bayesian sample size that satisfies a specific performance criterion can be simulated as follows: 1) generate parameters under the design prior; 2) given those parameters, generate a sample of size  $n$  according to the data density; 3) obtain the posterior distribution under the fitting prior; 4) from the posterior distribution, determine performance; 5) repeat steps 1–4 many times to obtain a distribution for performance; 6) from this distribution, determine if the performance criterion is met; 7) if the performance criterion is not met, increase the sample size to greater than  $n$ , and repeat steps 1–6. These steps are repeated until a sample size is reached for which the performance criterion is met. The Bayesian approach to sample size, as outlined above, can involve intensive simulation. For some situations, simpler methods can be appropriate to obtain an approximation to the sample size required.

To illustrate, consider the MACE rate parameter  $p_1$  in Section 2.1, for which the claim is  $p_1 < 0.249$ . The fitting prior might be the prior agreed upon by the FDA and the sponsor. The design prior could be the fitting prior conditional on MACE rate being less than 0.249. The decision rule is that the claim  $p_1 < 0.249$  is demonstrated if its posterior probability is greater than 0.975. A sample size can be determined such that if data are generated under the design prior, then the claim has power 80%, say, of being demonstrated with the posterior analysis under the fitting prior.

### 3.3. A Power Approach to Sample Size

The sample size needed to obtain sufficient power to detect a claim is sometimes relatively easy to compute. Consider again the cardiovascular device example. As in Section 2.3 we exploit that the posterior probability for MACE rate  $p_1 < 0.249$  is monotonically decreasing in event count  $y_1$  for the current study. Assuming a value in the alternative hypothesis space  $p_1 = p_A < 0.249$  (a degenerate design prior), the power at sample size  $n$  to achieve the success criterion (i.e.,  $p_1 < 0.249$  with posterior probability 0.975), can be computed in two steps: 1) find the largest count  $y_1^*$  such that the posterior probability (under the fitting prior) of  $p_1 < 0.249$  is greater than 0.975, and 2) compute power as  $Pow(n, p_A) = \Pr(y_1 \leq y_1^* | n, p_A)$ , the cumulative distribution of  $y_1$  evaluated at  $y_1^*$ , where  $y_1$  has binomial distribution  $Bin(n, p_A)$ . To illustrate, under binomial-normal hierarchical model as

expressed in Equation (1), the study success criterion is achieved if 44 or fewer MACE events are observed among the 200 patients in the study. Assuming  $p_A = 0.17$ , the power is then  $\Pr(y_1 \leq 44 | n = 200, p_A = 0.17) = 0.973$  or 97.3%. For the alternative hyperprior  $\sigma_\mu^{-2} \sim \Gamma(10, 10)$  discussed Section 2.1, the success criterion is achieved with 38 or fewer events, leading to a power of 80.3%.

This power approach can be extended to nondegenerate design priors. For a nondegenerate design prior  $\pi(p_1)$  restricted to values  $p_1 < 0.249$ , the average power can be computed in a pre-posterior analysis as  $Pow(n) = \int_P Pow(n, p_1)\pi(p_1)dp_1$  over the support  $P = \{p_1 : p_1 < 0.249\}$ .

### 3.4. A Hybrid Bayesian-Frequentist Approach to Bayesian Sample Size Determination

The frequentist approach can be leveraged in a novel preposterior analysis to determine a Bayesian sample size for a study with prior information. First, the frequentist sample size  $N_F$  is determined based on power and Type I error rate considerations. Next, given the values of the parameters assumed in the frequentist calculation (a degenerate design prior), data for a sample of size  $N_B$  are generated repeatedly and posterior analysis with the fitting prior is used on each dataset to obtain a distribution for the effective sample size  $ESS$ . The sample size  $N_B$  for the study is chosen as the smallest one such that the average  $ESS \geq N_F$ . From Equation (2),  $N_B$  is a sample size for which the Bayesian analysis will have about the same expected precision as the frequentist analysis that ignores the prior information. Therefore, the Bayesian analysis at sample size  $N_B$  would have approximately the same power as the frequentist analysis at sample size  $N_F$  if the Bayesian and frequentist decision rules are comparable in terms of level of evidence required for demonstrating a claim. Alternatively, the study could be monitored until the observed  $ESS \geq N_F$ . Monitoring of a study for  $ESS$  could be done so as not to reveal the observed effect size if revealing it could potentially compromise the integrity of study results.

## 4. LEVERAGING HISTORICAL CONTROLS

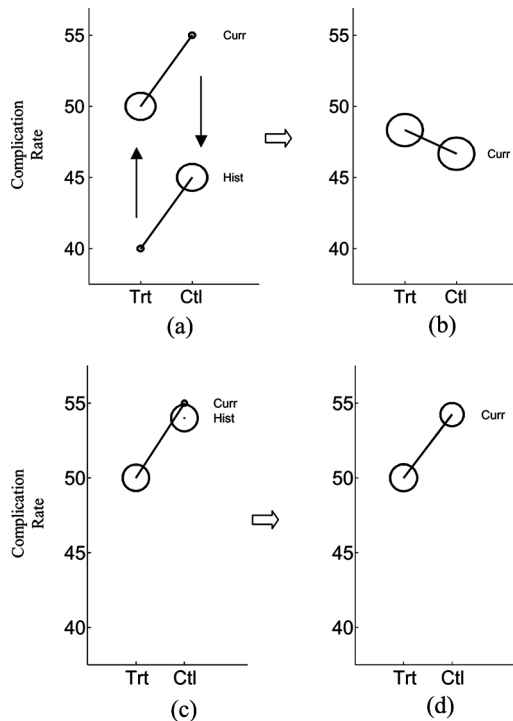
Historical controls are a common source of prior information for a randomized controlled study. Pocock (1976) presents excellent advice on the use of historical controls as prior information for randomized trials. Using hierarchical modeling, he derives a method for estimating the optimal number of subjects to randomize to treatment and control arms given a fixed total number of subjects. Generally, the optimal allocation is to randomize more subjects to treatment than to control because with equal allocation precision in the randomized control arm is larger due to the extra information utilized from the historical control(s). An important by-product is that, for the same sample size, the treatment under investigation can be studied in more subjects than if the historical control information were ignored.

Pocock bases his optimal allocation on minimizing the conditional posterior variance of the treatment effect given the variance for the treatment, the variance for the randomized control, and the variance between the historical and randomized

control studies. By specifying values for these variance components, the optimal allocation ratio of treatment to control subjects for a fixed total sample size is determined. This approach can be extended by applying a design prior on the variance components to obtain the optimal allocation ratio accounting for uncertainty in their value.

#### 4.1. Calibrating for Covariates and Exchangeability

When combining prior information with current study data, proper adjustment for all measured covariates can be crucial for a valid analysis. For example, consider a randomized controlled noninferiority study of a complication rate. Suppose a hierarchical model is applied to borrow strength from historical controls to help estimate the concurrent control rate so that the sample size of the concurrent control can be smaller than the treatment group [see Fig. 1(a); circle sizes are proportional to sample size]. If, for example, patients in the pivotal



**Figure 1** Illustration of adjustment for covariates (Circle sizes are proportional to effective sample size). In panels (a) and (b), borrowing within arms without adjusting for covariates can make inference about noninferiority unnecessarily more difficult because the concurrent control rate is pulled down toward the historical control rate, and the historical treatment rate is pulled up toward the current treatment rate. In panels (c) and (d), borrowing is done within the control arm after adjusting for health-related covariates. The adjustment leads to less separation in the concurrent and historical control complication rates, leading to the concurrent control rate borrowing from the historical control rate without the point estimate being affected substantially.

study (“Curr” in the figure) are more likely to have complications than those in the historical control studies (“Hist”) because they are less healthy, then borrowing strength from the historical controls without proper adjustment for covariates measuring health will bias downward the concurrent control complication rate [see Fig. 1(b)]. In this case, the bias will incorrectly make it more difficult to demonstrate that the device is noninferior to the control than if the historical control data were ignored. This phenomenon is less likely to occur if the analysis is adjusted for all measured covariates, which could be achieved via regression or propensity score methodology (Rosenbaum and Rubin, 1983). In Fig. 1(c), the concurrent control borrows from historical controls, after adjustment for covariates. The resulting analysis of the rates in Fig. 1(d) facilitates inference about noninferiority.

#### **4.2. Borrowing on Treatment-Control Differences Versus Borrowing Within Arms**

We have illustrated borrowing within treatment arms in Section 4.1. Borrowing across treatment-control differences and borrowing within treatment arms can make different assumptions, and have different effects. Borrowing within arms assumes the treatments are exchangeable, and separately, that the controls are exchangeable. Borrowing on treatment-control differences assumes that the study-specific differences between treatment and control are exchangeable. Often, we are forced to use the first method over the second because while a control group might be in every study, the treatment might only be in the latest study. However, while the study-specific treatment effects and study-specific control effects could vary substantially from study to study, the study-specific differences between treatment and control means might remain relatively constant, enabling the second method to achieve a much higher precision on the treatment-control difference. This method in effect is using the control as a baseline covariate to calibrate the studies, whereas borrowing within arms could require additional adjustment for baseline covariates.

We note that a concurrent or historical control group should not be used as a source of prior information for the treatment group, especially to show that the new device is noninferior to the control device. Having the quantitative prior information serve both as the prior and also as the comparator is circular and assumes that there is prior evidence of noninferiority.

#### **4.3. Synthetic Control for One-Arm Studies**

Medical device studies are sometimes one-armed, relying on historical controls as comparators rather than randomized concurrent ones. For example, a cardiovascular device studied in a single-armed trial may be compared with a rather antiquated device in historical controls, say a procedure to which no one would assign patients now. Because the gold standard in clinical trials is comparison with a concurrent randomized control, if possible, the endpoint for statistical analysis should try to match this comparison as closely as possible. One approach is to compare the treatment group in a one-arm study with a synthetic control, a new control group derived from historical control studies that can be regarded as the concurrent control group for the treatment.

The idea behind deriving the synthetic control is to model two or more historical control population means as exchangeable with the mean of a synthetic control. The synthetic control is an unobserved random draw from the exchangeable distribution on which no data have been collected. However, because the synthetic control is related to the historical controls via exchangeability, its posterior distribution given the historical control data is informative. The goal of the statistical analysis is to compare the treatment population mean with the synthetic control population mean.

To fix ideas, suppose that we would like to compare a new device for coronary plaque reduction to a concurrent control procedure with respect to percent diameter stenosis (%DS), and we would like the difference to be less than zero, with posterior probability at least 0.95. But the study does not contain a concurrent control group. However, there are two historical control groups from two previous studies available. We might model these studies as exchangeable with a hypothetical (synthetic) control group, with respect to %DS.

In order to compare the treatment rate with the synthetic control rate, we need to ensure that we measure enough of the important covariates to make a fair comparison between the new treatment group and control group rate drawn from the population of control group rates. To assess whether there is a study by covariate interaction, these covariates need to be measured in each study at the level of the patient. They cannot be merely summaries, such the covariate means one might obtain from a report of a historical study in the literature.

As an example of an important covariate, consider that a diabetic has greater risk of heart disease. Suppose that the historical control groups end up differing greatly from the treatment group in the percent of diabetic patients. So, whether or not a patient is diabetic is a patient-level covariate that must be included in the model before the studies can be considered exchangeable. As a simple illustration, data are given in Table 4 for 10 patients per group.

These data show that the mean %DS is 0.50 for the new device, and 0.48 and 0.45 for the two historical control groups. However, the new device group ended up having many more diabetic patients than the control groups. So, the reason for the higher mean may very well be because of the greater number of diabetics (those with

**Table 4** Data from a one-armed study and two historical controls

New device		Historical control 1		Historical control 2	
%DS	Diabetic	%DS	Diabetic	%DS	Diabetic
0.40	0	0.40	0	0.35	0
0.40	1	0.40	0	0.35	0
0.45	1	0.44	0	0.40	0
0.45	1	0.44	0	0.40	0
0.50	1	0.48	0	0.45	0
0.50	1	0.48	0	0.45	0
0.52	1	0.52	0	0.50	0
0.52	1	0.52	1	0.50	0
0.53	1	0.56	1	0.55	1
0.53	1	0.56	1	0.55	1

**Table 5** Posterior estimates of parameters for synthetic control example

Parameter	Mean	SD	2.5%	Med	97.5%
$\beta$	10.4	2.2	6.1	10.3	14.8
$\alpha_1$	38.6	2.5	33.5	38.6	43.4
$\alpha_2$	44.4	1.4	41.4	44.4	47.0
$\alpha_3$	43.6	1.6	40.2	43.7	46.6
$\alpha_{syn}$	42.9	14.1	26.2	44.0	52.3
$\alpha_0$	43.2	5.7	31.5	43.9	48.5
$\alpha_1 - \alpha_{syn}$	-4.4	14.3	-14.3	-5.3	13.0
$\Pr(\alpha_1 - \alpha_{syn} < 0)$	0.91	0.28	0.00	1.00	1.00

diabetes have the highest %DS). We account for this covariate in the model because we want to measure device effectiveness, not effectiveness due to not being diabetic.

Let  $y_{ij}$  be the %DS for group  $t$ , and patient  $j$ . For simplicity, we assume that %DS is normally distributed, with mean dependent on treatment group and whether the patient is diabetic ( $x_{ij} = 1$ , if diabetic and 0 otherwise). The model is then

$$y_{ij} \sim N(\alpha_t + x_{ij}\beta, \sigma_t^2)$$

for  $j = 1, \dots, n_t$ . We place normal priors on all parameters so that the treatment mean for the new device is considered a fixed effect, and the means for the controls are random effects, coming from a common distribution. We complete the modeling with noninformative hyperpriors.

$$\begin{aligned} \alpha_1 &\sim N(0, 1000) \quad (\text{new device}) \\ \alpha_2, \dots, \alpha_T &\sim N(\alpha_0, \sigma_\alpha^2) \quad (\text{historical controls}) \\ \alpha_{syn} &\sim N(\alpha_0, \sigma_\alpha^2) \quad (\text{synthetic control}) \\ \alpha_0, \beta &\sim N(0, 1000), \quad \sigma_t^{-2}, \sigma_\alpha^{-2} \sim \Gamma(.001, .001) \end{aligned}$$

Note that drawing the synthetic control mean,  $\alpha_{syn}$ , from the population of control means accounts for any remaining unobserved covariates through the size of the variance parameter. If there is still a lot of variability around the regression line after accounting for measured covariates, then the variance parameter will have a posterior distribution that is skewed toward higher values.

The parameter of interest is  $\alpha_1 - \alpha_{syn}$ . The posterior probability that this difference is less than zero is 0.91, which is less than the criterion of 0.95 (see Table 5). So, the study was not quite successful. The posterior variability of  $\alpha_{syn}$  is large because it is drawn from a population fit with only two historical controls. This causes the credible interval on  $\alpha_1 - \alpha_{syn}$  to be large. Measuring additional covariates could reduce the posterior variability. This example used only one covariate for illustrative purposes.

## 5. SUBGROUP ANALYSIS

Subgroup analysis is a frequent issue in the evaluation of medical device trials under regulatory review. It is generally advisable that subgroup analysis be planned prospectively, because otherwise subgroup findings would be considered merely

exploratory. A perspective on subgroup analysis for medical device clinical trials is given in Scott and Campbell (1998). A common Bayesian approach to subgroup analysis is to adjust the inference on a subgroup by borrowing strength from other subgroups thought to be related to it. Typically, a set of subgroups is identified as having effects that are exchangeable in their prior distribution. These effects are assumed to be drawn randomly from a common distribution. This random effects (hierarchical) model relates the subgroup effects because information on one effect provides information on the common distribution, which in turn provides information on the other subgroup effects. This approach has been proposed for clinical trials applications by Dixon and Simon (1991, 1992).

### 5.1. A Bayesian Subgroup Analysis

To illustrate with a pharmaceutical example, consider the Losartan Intervention For Endpoint reduction in hypertension (LIFE) study, a multinational, double-blind study comparing Losartan (brand name COZAAR) and atenolol in 9193 hypertensive patients with ECG documented left ventricular hypertrophy (<http://www.fda.gov/cder/foi/label/2003/020386s0321bl.pdf>). The primary endpoint was first occurrence of cardiovascular death, nonfatal stroke, or nonfatal myocardial infarction. A Cox analysis stratified by race indicates that losartan was significantly better than atenolol in nonblacks ( $p = 0.003$ ) but significantly worse in blacks ( $p = 0.033$ ) (Table 6). A question is whether the multiple racial subgroups generated a directional error in blacks (losartan is better than atenolol in blacks, but the opposite was concluded).

A Bayesian subgroup analysis could be conducted by assuming the maximum likelihood estimates (MLEs) of the log hazard ratios for blacks and nonblacks are normally distributed about their true values with variances equal to their standard errors. The true values are then assumed exchangeable, drawn independently from a common normal distribution with a mean and variance that are given diffuse hyperpriors. The model is

$$y_i \sim N(\mu_i, s_i^2), \quad \mu_i \sim N(\mu_0, \sigma_\mu^2), \quad \mu_0 \sim N(0, 1000), \quad \sigma_\mu^{-2} \sim \Gamma(0.001, 0.001)$$

**Table 6** Cox and Bayesian analysis of LIFE study stratified by nonblack and black racial subgroups

	log HR	SE	HR (95%CI <sup>†</sup> )	P
Nonblack				
Cox analysis	-0.19	0.06	0.83 (0.73, 0.94)	0.003
Bayesian	-0.18		0.84 (0.74, 0.94)	
Black				
Cox analysis	0.51	0.24	1.67 (1.04, 2.67)	0.033
Bayesian	0.38		1.47 (0.87, 2.44)	

<sup>†</sup>95% CI refers to the 95% confidence interval for the Cox analysis and a 95% central posterior interval for the Bayesian analysis.

where for the two race groups  $i = 1, 2$ ,  $y_i$  is the Cox analysis MLE of the log hazard ratio and  $s_i^2$  is its standard error,  $\mu_i$  is the true log hazards ratio, and  $\mu_0$  and  $\sigma_\mu^2$  are the mean and variance of the normal distribution assumed for the true log hazards ratios.

In contrast with the standard Cox analysis, the Bayesian analysis indicates for blacks a smaller, insignificant hazard ratio comparing losartan with atenolol (95% central posterior interval contains one) (Table 6). For nonblacks, the hazard ratio was nearly the same and remained significant (95% central posterior interval entirely below one). We emphasize that the Bayesian analysis is predicated on a priori exchangeability of nonblacks with blacks in the hazard ratio, in other words, on not expecting a smaller effect in blacks than in nonblacks a priori.

## 5.2. Bayesian Analysis of a Confirmatory Study for a Subgroup

When a trial fails on the primary endpoint, a post hoc subgroup analysis is sometimes made in an attempt to find promising subgroups. In this case, a confirmatory trial is needed for any promising subgroup found, because proper adjustment for post hoc multiple subgroup testing is difficult, if not impossible. The problem is that without a prospective plan for subgroup analysis, the subgroups that were submitted by the sponsor could be a small subset of the subgroups that were actually considered, and that this larger set is unknown.

If the confirmatory study is frequentist, then the information that was gathered on that subgroup from the failed study is not used. A Bayesian confirmatory study can consider that information as prior information. However, particular care needs to be taken in developing the prior distribution. If the sponsor considered a great number of subgroups, the promising effectiveness or safety results observed in the subgroup could easily be a spurious finding. One approach to developing the prior distribution is to first apply a post hoc Bayesian adjustment to the subgroup results from the failed study by structuring the subgroup as exchangeable with other subgroups considered in that study.

To illustrate, suppose a randomized controlled trial (Study 1) is conducted to show that a therapeutic device is more effective than a control treatment in reducing infarct size for acute myocardial infarction (MI), but the trial failed to reach statistical significance. However, a post hoc analysis shows that statistical significance was met within a particular subgroup (Group 1) of patients for whom the interval between MI onset and time to reperfusion was less than a certain number of hours. Thus, we run a new trial (Study 2), only enrolling patients who fall in this subgroup, and randomizing them to device or control.

To develop a prior distribution for the subgroup effect in Study 2, suppose that the remaining patients in the first trial can be divided into three additional subgroups (Groups 2, 3, and 4), such that all four subgroups can be considered exchangeable in the treatment effect on infarct size. Note Study 1 contains all four subgroups, but Study 2 contains only one subgroup (Group 1), the one of interest. A prior distribution for the subgroup effect in Study 2 is obtained by modeling that the subgroups in Study 1 are exchangeable in the treatment effect. This prior distribution is used in Study 2 by additionally assuming that the studies are exchangeable in the treatment effect for Group 1. These assumptions are implemented with a hierarchical model.

The entire model for some transformation of infarct size,  $y$ , can be written as

$$E(y_{sgt}) = \mu + \xi_s + \gamma_g + \tau_t + (\xi\tau)_{st} + (\gamma\tau)_{gt}$$

$$\xi_s \sim N(0, \sigma_s^2), \quad \gamma_g \sim N(0, \sigma_g^2), \quad (\xi\tau)_{st} \sim N(0, \sigma_{st}^2), \quad (\gamma\tau)_{gt} \sim N(0, \sigma_{gt}^2)$$

Here  $E(y_{sgt})$  is the mean response for patients in study  $s$  ( $s = 1$  for failed Study 1,  $s = 2$  for confirmatory Study 2), in group  $g$  ( $g = 1, 2, 3, 4$ ), and on treatment  $t$  ( $t = 1$  for device,  $t = 2$  for control). Exchangeability of the subgroup-specific treatment effects in Study 1 is reflected by a random subgroup by treatment interaction effect  $(\gamma\tau)_{gt}$ . Exchangeability of the study-specific treatment effects for Group 1 is reflected by a random study by treatment interaction effect  $(\xi\tau)_{st}$ . Main study effects ( $\xi_s$ ) and subgroup effects ( $\gamma_g$ ) are also assumed to be random, while main treatment effects ( $\tau_t$ ) are assumed fixed. The quantity of interest for validating device effectiveness in Group 1 is  $E(y_{212}) - E(y_{211})$ , the treatment effect in the confirmatory study (Study 2) for Group 1. The model is completed by placing diffuse hyperpriors on the variance components, the treatment effects, and the grand mean  $\mu$ .

To illustrate, we generate hypothetical data on infarct size for ten subjects per treatment group in the four subgroups in Study 1 and the one subgroup in Study 2, for 100 subjects. In Study 1, the overall treatment effect was insignificant ( $z = 0.70$ , 2-sided  $p$ -value = 0.487), but for Group 1, the subgroup of interest, the treatment effect was significant ( $z = 2.40$ , 2-sided  $p$  value = 0.016) (Table 7). In Study 2, the treatment effect in Group 1 was in the right direction, but nonsignificant ( $z = 1.95$ , 2-sided  $p$ -value = 0.051). The Bayesian analysis shrunk the estimate of the Group 1 treatment effect in Study 1 (posterior mean < sample mean) and slightly increased its precision (posterior standard deviation < sample standard error), resulting in the finding still being significant (95% credible interval above 0), but less so. For the Group 1 treatment effect in Study 2, the Bayesian analysis slightly increased the estimate (posterior mean > sample mean) and maintained the precision (posterior standard deviation = sample standard error), resulting in the insignificant finding becoming significant.

The effective sample size was nearly the same for Study 2 than if the prior information was ignored ( $ESS = 100 * (0.0218/0.02157)^2 = 102.1$  versus  $N = 100$ ). A reason the uncertainty was not decreased is the imprecision with which the between study variance  $\sigma_s^2$  and the study by treatment interaction variance  $\sigma_{st}^2$  are

**Table 7** Parameter estimates for the subgroup example

Study	Group	Control	Device	Difference	Z	Posterior analysis	
		Mean (SE)	Mean (SE)	Mean (SE)		Mean (SD)	95% CI
2	1	0.254 (0.46)	0.212 (0.051)	-0.042 (0.022)	-1.952	-0.043 (0.022)	(-0.085, -0.0)
1	1	0.253 (0.039)	0.201 (0.056)	-0.052 (0.022)	-2.404	-0.044 (0.021)	(-0.085, 0.004)
1	2	0.199 (0.041)	0.226 (0.050)	0.027 (0.020)	1.348	0.021 (0.021)	(-0.020, 0.062)
1	3	0.234 (0.056)	0.227 (0.069)	-0.006 (0.028)	-0.225	-0.007 (0.021)	(-0.048, 0.034)
1	4	0.192 (0.034)	0.234 (0.036)	0.042 (0.016)	2.723	0.036 (0.021)	(-0.005, 0.078)
1	All	0.227 (0.051)	0.220 (0.053)	-0.007 (0.010)	-0.695		

estimated, because there are only two studies and two treatments. In such cases, the analysis is sensitive to the prior distribution for the variance components. We settled on giving the inverse of each of the variances (their precisions) a gamma hyperprior distribution with parameters 1 and 0.0025 to center the precisions at 400, the observed standard deviation of individual infarct sizes. Other prior distributions would give different results. For a discussion of alternative, uniform prior distributions for variance component standard deviations, see Gelman (2006).

In the WinBUGS code for the Bayesian analysis of this example (Appendix II), a centering parameterization (Gelfand et al., 1995) was used to facilitate better mixing and therefore a more rapid convergence of the Markov chain of sampled parameter values to the target distribution.

## 6. BAYESIAN MONITORING AND ADAPTIVE DESIGN

### 6.1. Monitoring

Frequently, companies wish to monitor their trials by performing interim analyses during the study period. At an interim time point, the company may make a decision to stop enrollment or to stop follow-up of patients, based on a statistical stopping rule. The stopping rule might also include a decision to declare the device effective or to declare the trial futile in determining effectiveness, given the value of a posterior criterion. In the frequentist world, these trials are called group sequential trials (see Jennison and Turnbull, 2000). A frequentist group sequential trial requires adjustment of the usual critical value to declare significance at an interim time point, because of an increase in type I error rate caused by multiple looks at the data. Because Bayesian inference is on the parameter space and not on the data space (which is considered fixed), in theory, a Bayesian design with interim monitoring needs no adjustment of the criterion for trial success. However, if one desires to control the frequentist type I error rate of the design, then the criterion might need to be made more stringent relative to a single stage design.

We have seen two general strategies for Bayesian monitoring: 1) Stop the trial early at a given stage if the posterior probability of the primary hypothesis of interest is larger than a fixed threshold, and 2) Calculate at each interim stage the predictive probability that the hypothesis will be demonstrated at the end of the study once the data on all subjects have been obtained, and stop early if the predictive probability is larger than a threshold. If, respectively, the posterior or predictive probability is very low, then the trial could also be stopped for futility. Whereas Strategy 1 tends to inflate overall frequentist type I error rate due to multiple looks and decisions, Strategy 2 tends to be more conservative at interim stages because it accounts for uncertainty in the unknown end-of-trial values. Within either of these strategies, one can use interim looks to determine the total sample size. For example, if the predictive probability of trial success is greater than a given threshold, one might stop the accrual of patients, and call the current number the total sample size. After halting accrual, one can then conduct further interim analyses by monitoring patient follow-up to determine whether to stop the trial early for success.

As an example, consider an embolic protection device designed to reduce the stroke rate caused by embolic material accidentally released after stenting the

carotid artery. Suppose this device is proposed to be noninferior to endarterectomy (a more invasive surgical excision of plaque) with respect to stroke or other adverse events at one-year follow-up, with a noninferiority margin of 2%. A maximum of 1500 patients will be enrolled and randomized 2:1 to device or surgery. If  $\theta_T$  denotes the event rate in the device group, and  $\theta_C$  denotes the event rate in the endarterectomy group, then the device is considered noninferior to endarterectomy if the posterior probability that  $\theta_T - \theta_C < 2\%$ , is greater than 0.97.

In this trial there is no objective prior information. We use a noninformative prior to do monitoring for possible early stopping (a monitoring prior) and for final comparison of the event rates in the two groups (a fitting prior). However, when monitoring the trial using predictive probabilities, we might alternatively use an informative prior centered near “enthusiastic” results for the device (sometimes called an “enthusiastic prior”) if we are very confident that the device will outperform the control, and we will not need all 1500 patients in order to show noninferiority. In this way, we can ensure a good chance of stopping before enrolling 1500 patients. For the final analysis once all enrolled patients complete one-year follow-up, we would like to use a noninformative (fitting) prior in order not to bias results with subjective opinion. However, one must be especially careful with using an enthusiastic prior during monitoring so as not to stop enrollment too early. We return to this warning after the example.

Suppose the device is actually slightly superior to surgery, with an event rate of 4.5%, whereas surgery has an event rate of 5%. These rates are chosen strictly for illustrative purposes, and do not necessarily correspond to actual rates. For simplicity, we assume these are Bernoulli proportions. With interim looks at 500 and 1000 patients, suppose we halt accrual at 1000 patients, after obtaining a predictive probability of 0.996 of a successful complete trial. This predictive probability was computed with uniform priors on the event rates (i.e.,  $\beta[1, 1]$ ).

At the time that enrollment stopped, suppose only 750 patients out of the 1000 enrolled had completed the one-year primary endpoint, with a 4.2% observed event rate in the device group, and a 6.8% observed event rate in the surgery group. At this point the trial is then monitored for possibly stopping for early success (or early futility). Suppose we plan to conduct two interim analyses; one when 125 additional patients have completed follow-up or had an event, and the other when all 1000 patients have completed follow-up or had an event. The priors on the event rates will be  $\beta(1, 1)$  priors. At the first interim analysis, the posterior predictive probability of success at the end of the trial is 0.998, which exceeds the criterion of 0.97. This predictive probability was obtained with 583 device patients and 292 control patients (with 3.8% and 6.2% observed event rates), much fewer than the original maximum sample size of 1500 patients. Incidentally, the maximum sample size gives about 70% frequentist power to conclude noninferiority if the true device rate is less than the true control rate by 0.5%. Also, the frequentist type I error rate is approximately 0.06 if we assume that the true device rate is 0.02 greater than the control rate, and then simulate trials using this assumption. Frequentist operating characteristics and average total sample size, assuming control rates of 4 and 5%, after 1000 simulations is given in Table 8. We see that the chosen posterior probability criteria enjoy good frequentist properties.

With monitoring for enrollment, we only needed to enroll 1000 patients in order to have a high predictive probability of a successful study. The device was

**Table 8** Results from 1000 simulations of Bayesian monitoring of stroke rate trial

Assumed true rates		Average total sample size	Estimated power/type I error rate
$\theta_C = 0.05$	$\theta_T = 0.04$	1016	0.83
	0.045	1088	0.75
	0.05	1223	0.54
	0.07	1469	0.06
$\theta_C = 0.04$	$\theta_T = 0.03$	877	0.89
	0.035	1044	0.84
	0.04	1221	0.55
	0.06	1463	0.06

actually superior to the surgery here (if it were equivalent to surgery, then results are very similar). Thus, the trial could have been monitored using an enthusiastic prior for the device performance. However, in this case, if enrollment is stopped at 500 patients, the success criterion is not met for the final analysis (posterior probability = 0.963). From Table 8, we see that the average total sample sizes needed are well above 500 even when superiority holds. So, one must be careful when using enthusiastic monitoring priors not to stop too early, unless it is certain that the device is more than marginally better than control. Simulations such as those above can help to determine the risk of stopping too early by using monitoring priors that are too enthusiastic.

## 6.2. Adaptive Design

Besides adaptive sample size designs, another type of adaptive design is to consider dropping (or adding) an arm midway through a trial. This type of adaptive allocation might be done when a particular active control or placebo is thought to be potentially harmful or undesirable to patients, or a newly approved device comes onto the market and could be used to replace the placebo. Berry (2004) reviews Bayesian methodology for handling adaptations such as these without compromising the integrity of trial conclusions. Essentially, as information accumulates in the trial (and outside of the trial), the predictive probability of a treatment's success is used to adjust its "weight" for the adaptive allocation. Berry concludes that the adaptive allocation results in better treatment for patients in the trial because they eventually begin receiving the best performing treatment.

Although Berry focuses mostly on drug trials and phase II to III transitions, adaptive allocation could be done for device clinical trials as well. For example, the relatively rapid development of devices can make a placebo arm in a device trial essentially obsolete. During the planning stages of a trial, a competing device might be approved as superior to placebo, possibly necessitating a switch during the trial from randomizing to placebo to randomizing to the newly approved device once it gains market penetrance at the sites at the which the trial will be conducted. However, only showing noninferiority to the newly approved active control might not be sufficient evidence that the device is more effective than placebo. Thus, a comparison to placebo is needed. But, it could be potentially unethical to randomize

to placebo if the trial began to show evidence of superiority rather quickly. An alternative is to use adaptive allocation with three arms, a placebo, an active control, and the experimental device. This allows elimination of the placebo arm if the predictive probability of success is lower than a particular number.

Other types of adaptive modifications can be done, such as changing inclusion/exclusion criteria, changing performance criteria, or adding centers. Although it is difficult to anticipate the need for such changes prospectively, planning for possible changes can greatly facilitate inference at the end of the trial and maintain trial integrity.

## 7. MODEL CHECKING AND SENSITIVITY ANALYSIS

### 7.1. Model Checking

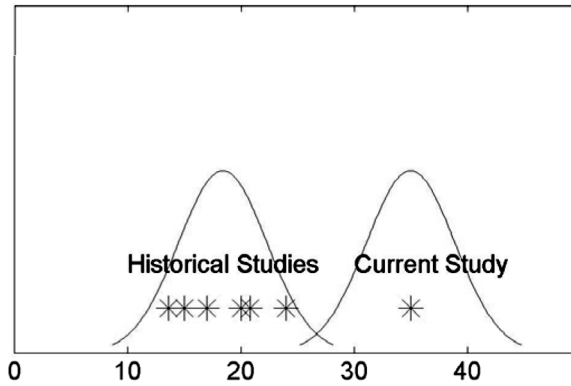
Model checking is important in any analysis, but tends to be especially important in a Bayesian analysis due to the model complexity that is often permitted in these analyses. Examples of model checking tasks include checking assumptions made in a Bayesian analysis such as exchangeability of parameters, and conducting sensitivity analyses to see how robust study conclusions are to changes in the prior, in the modeling, or in how missing data are imputed.

A tool frequently used in model checking is the posterior predictive distribution of future, hypothetical observations from the trial, under the assumptions of the current model. The posterior predictive distribution of a new observation,  $y_{new}$ , from model  $p(y|\theta)$ , when  $y_{old}$  are data already observed is  $p(y_{new}|y_{old}) = \int p(y_{new}|\theta)p(\theta|y_{old})d\theta$  (Gelman et al., 2003).

One can generate new, hypothetical data from the posterior predictive distribution and see how well it “matches” observed data from the trial. The assessment of how well the posterior predictive data match the observed data can be made using the “Bayesian  $p$ -value”, which is the posterior predictive probability of observing a result at least as extreme as what was observed in the trial (see Gelman et al., 1996, 2003).

For example, suppose we want to check the assumption of exchangeability. Exchangeability is determined from a clinical and engineering standpoint at the planning stage. However, we can also check this assumption statistically. Suppose the hypothetical trial described in Table 1 yielded a MACE rate of 0.35. This rate is much higher than the rates from the historical studies, despite a clinical judgment of exchangeability of studies prior to collecting data on the new study. If there are no observed covariates that could account for the difference, we might use posterior predictive analysis to check whether the new study rate is likely to have come from a different distribution than the historical rates.

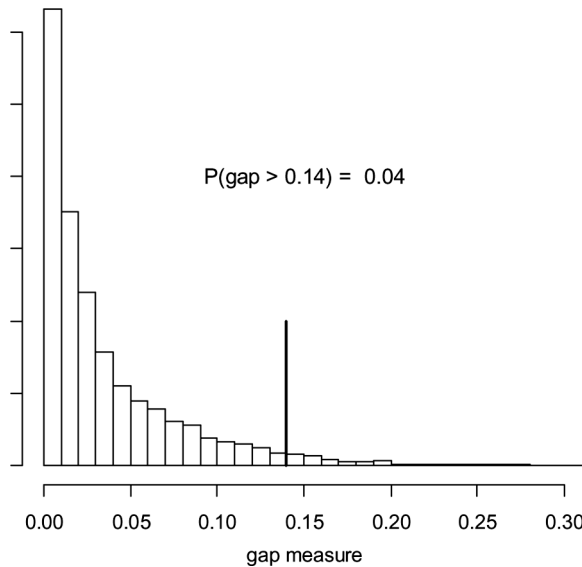
Figure 2 plots the observed MACE proportions for each of the six historical studies from Table 1, and for the hypothetical current study mean. If the between-study variability is much smaller than the maximum difference in observed rates, then the likelihood of exchangeability is low. i.e., the exchangeable model does not fit the data well, and we might have two separate distributions, as shown in Fig. 2. We can quantify this likelihood with the posterior predictive distribution of a newly observed statistic, chosen to be particularly sensitive to departures from exchangeability. The posterior predictive distribution is obtained by drawing



**Figure 2** The smoothed distributions drawn over the observed values depict two separate possible distributions, instead of one exchangeable distribution.

hypothetical new data given the observed data, assuming exchangeability of studies. The test statistic we chose is the smallest absolute difference between the proportion for the new (seventh) study and another study proportion, a type of gap statistic. If the observed gap is located far into the right-hand tail of its posterior predictive distribution, then the exchangeable model may not hold.

The posterior predictive distribution of the gap statistic using observed proportions drawn from seven new hypothetical studies consistent with an exchangeable model is given in Fig. 3. The observed value of the test statistic is 0.14, and is indicated with a dark vertical line. The probability of observing a gap value,



**Figure 3** Posterior predictive distribution of the “gap” statistic (see text). The dark line indicates the observed gap statistic of 0.14.

from the exchangeable model, that is at least as large as that observed in the data is about 0.04. This might be considered small enough to question exchangeability.

A well-known potential problem with using the posterior predictive distribution for model checking is that because the new, hypothetical data are not marginally independent of the data used to fit the model; there is a double use of the data for fitting, as well as for checking the fit. Alternatives for model checking include having separate fitting and testing samples (though both samples should ideally not be collected simultaneously) or using cross-validation. The cross-validated Bayesian residual is computed using estimated moments from the conditional predictive distribution: the distribution of a data point, conditional on the remaining points (see Carlin and Louis, 2000, p. 47).

In addition, Bayesian deviance measures such as Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) can be used for model choice by comparing the fit of one model over another. DIC appears as an option in the WinBUGS program. It is calculated as the posterior mean of the model deviance minus the effective number of parameters in the model ( $p_D$ ), where the effective number of parameters is the decrease in deviance expected from estimating the parameters of the model.  $p_D$  can be estimated as the posterior mean of the deviance minus the deviance evaluated at the posterior mean of the parameters (Gelman et al., 2003, p. 182, give a good discussion of effective number of parameters). The minimum DIC estimates the model that will make the best short-term predictions. A useful application of DIC is to compare the fit of a normal random effects distribution to that of a (truncated) Dirichlet process mixture (DPM) model, if a more flexible random effects distribution is desired (Ohlssen et al., 2007). We discuss this type of application briefly in Section 9.

## 7.2. Sensitivity Analysis

Inference for complex Bayesian models involving hyperparameters can sometimes be sensitive to the prior distribution on the hyperparameters (hyperpriors). Sensitivity to the prior distribution can arise, for example, if the data are not very informative for the hyperparameters. In earlier sections, we give a variety of examples in which prior distributions are placed on hyperparameters such as variance components. While these examples are illustrative, we caution that the results given depend on the choice of hyperprior. This can present a challenge in determining the best inference for a given dataset. As a matter of course, sensitivity of inferences to the prior distribution should be checked in any Bayesian analysis.

An important example of sensitivity analysis is to check robustness of inference to departures from a missing data mechanism assumed in the Bayesian analysis. To illustrate, suppose that for the example in Section 6, where the stroke rate for device was noninferior to surgery, the complication rate was also measured at both 12 and 24-months follow-up, and we want to try to claim noninferiority of the long-term complication rate at 24 months. While patients who have not yet reached 24-months follow-up might be safely assumed to be missing at random, patients who missed their 24-month visit may exhibit informative missingness, that is, may have missed the visit for reasons related to their 12- or 24-month outcome.

Suppose that 100 patients were lost to follow-up for the 24-month measurement. Suppose that the posterior probability of noninferiority of long-term

complication rate exceeds the prespecified criterion, but that for patients lost to 24-month follow-up, the 12-month complication rate tended to be high more often for the device group than for the control group. A Bayesian model could be constructed to adjust for this trend, where the 24-month event rate is dependent on the 12-month event rate, as well as any measured covariates. The Bayesian model would effectively impute the 24-month outcomes lost-to-follow up based on 12-month outcomes and account for the uncertainty in this imputation.

We might also consider scenarios where the missing device patients are more likely to have higher 24-month complication rates than the missing control patients, beyond what is predicted by the 12-month outcome (so-called “worse-case” scenarios). The observed fit of the relationship between 12 and 24-month outcomes can be perturbed differentially for device and control patients lost to 24-month follow-up until the noninferiority conclusion no longer holds. A statement could then be made about the robustness of the noninferiority conclusion to 24-month informative missingness. A real life situation where this type of sensitivity analysis was submitted to FDA in an approved PMA is the BAK/Cervical Interbody Fusion System by Sulzer Spine-Tech (2000, website [www.fda.gov/cdrh/pdf/p980048b.pdf](http://www.fda.gov/cdrh/pdf/p980048b.pdf)).

## 8. BAYESIAN METHODS FOR DIAGNOSTIC DEVICES

Bayesian studies have been less frequent for diagnostic devices than therapeutic devices. One possible reason is that for in vitro diagnostics, sampling can be cheap, mitigating the need for prior information. Therefore, the discussion in this section is in parts speculative. However, Bayesian methods for diagnostic medicine are given a comprehensive treatment in the recent book by Broemeling (2007).

A diagnostic device that was approved by FDA based on a Bayesian analysis is the Transcan T-2000 electrical impedance breast scanner device to diagnosing breast cancer ([www.fda.gov/cdrh/pdf/p970033b.pdf](http://www.fda.gov/cdrh/pdf/p970033b.pdf)). To estimate sensitivity and specificity in a small intended use study, a Bayesian multinomial-normal hierarchical model was used borrow strength from three other larger studies in which the device was not used exactly as intended. The Bayesian analysis indicated that sensitivity and specificity in the small intended use study was decidedly greater for the Transcan device used adjunctively with mammogram reading than for mammogram reading alone and that the use of Transcan would result in fewer biopsies. These findings would not have been conclusive were it not for the prior information from the three previous studies.

A common situation that occurs in trials of diagnostics for detecting a disease condition of interest is that the disease is verified in only a fraction of the subjects. The reason is that disease verification can be expensive. A typical scenario is that all the test positives are verified for disease, but only a fraction of the test negatives are. Because of this differential in verification, the ordinary estimates of sensitivity and specificity of the test are biased. However, Bayes Theorem can be used to obtain unbiased estimates (Greenes and Begg, 1985). When disease status is missing at random conditional on test result, Bayesian methods can in principle be used to adjust for the missing data, even when all other subjects with that test result are verified to have the same disease status (e.g., all verified test negative subjects are disease negative), a situation in which standard imputation methods fail.

Unfortunately, all too frequently, studies of diagnostic tests are devoid of a gold standard, i.e., a procedure that can determine the true diagnosis for a subject. Without a gold standard, accuracy parameters of a test (e.g., sensitivity and specificity) are unidentifiable under classical analysis. Recently, Bayesian methods have been developed that enable estimation of the sensitivity and specificity of two or more diagnostic tests and the prevalence of the condition being diagnosed by using prior information on them. (Beiden et al., 2000a; Black and Craig, 2002; Dendukuri and Joseph, 2001; Georgiadis et al., 2003). In these methods, the true diagnosis is modeled as a latent variable with a prior distribution placed on the hyperparameters of the latent variable distribution. The first methods developed for this problem focused on models that assume that multiple diagnostic tests are independent conditional on the true diagnosis. This assumption can make accuracy parameters identifiable under certain conditions but is often unrealistic. Unfortunately, methods that allow multiple tests to be conditionally dependent can lack robustness to model assumptions (Albert and Dodd, 2004). However, reliable and precise prior information on some of the tests and the prevalence could identify plausible sets of parameter values (Kosinski and Barnhart, 2003).

For monitoring devices for which the true value of the measurement being made is unknown, methods that are essentially Bayesian have been developed that enable comparison of the device measurement to the true value by modeling the latter as a latent variable. The methods require repeated measurements of the true value by two or more devices. Kupinski et al. (2002) and Hoppin et al. (2002) estimate the mean square error assuming that the measurements from the devices are independent conditional on the true value. Pennello (2003a) explores extensions to the estimation methods that do not assume conditional independence, with applications to pulse oximeters.

Bayesian methods could also be helpful for obtaining an exact analysis when the design of a diagnostic study is complex. For example, a common design when comparing a digital mammography system with standard film is a multi-reader multicase design, in which every reader reads every case with both modalities. With two breasts per woman, and two mammographic projections per breast to review, the cranio-caudal and mediolateral oblique views, the data exhibit a complex multivariate structure. Because both readers and cases are considered as random samples from their respective populations, statistical models that have been used include random effects for reader, case, case by reader interactions, and interactions of case and reader with modality (Beiden et al., 2000b). Bayesian methods can provide an exact analysis for this complex modeling (cf. Johnson and Johnson, 2006), whereas asymptotic frequentist methods may yield a poor approximation for the inference. Conventional bootstrapping of readers and cases may also fail because both are random samples and typically the number of readers (5–10) is too small to bootstrap.

## 9. CONCLUDING REMARKS

We have provided examples of Bayesian modeling, analysis, design assessment, and sample size determination for device trial applications selected for their variety and frequency of appearance. In complex settings, we have found that the Bayesian models presented have been useful for obtaining reliable inferences by systematically

accounting for multiple sources of uncertainty. We conclude with a few brief remarks on topics not yet covered.

Hierarchical models have been presented as a flexible approach to account for uncertainty in the use of the prior information for a study. However, other methods besides hierarchical modeling could be used when there is a concern that the hierarchical model may not discount the prior information enough when warranted or when a parametric distribution for the between-study effects is questionable. For example, in Section 7.1 we mentioned using a Dirichlet process prior to place a distribution on the possible distributions of study-specific parameters, perhaps with an exchangeable distribution as its mean. For the data described in that section, we compared the fit of several Dirichlet process mixture (DPM) models in addition to the normal exchangeable model in order to check whether a more flexible random effects distribution would better accommodate the “outlying” MACE rate of 0.35. DIC was used as the criterion. Although a DPM model in fact reached the minimum DIC, results among various DPM models and the normal exchangeable model were somewhat equivocal, and therefore details of the model comparison are excluded. The interested reader can obtain WinBUGS code from the authors upon request. We also note the existence of an R package for doing Bayesian nonparametric analyses, in particular those using Dirichlet process priors. The package, called DPpackage by Alejandro Jara Vallejos, can be obtained from CRAN (<http://cran.stat.ucla.edu/>).

A power prior (Ibrahim and Chen, 2000) provides another possible approach to discounting prior information in which the prior distribution is exponentiated by an amount that depends on the discrepancy between the prior information and the study data, with exponents of zero and one being the extreme cases in which the prior data are ignored, and the prior data are considered exchangeable with the study data, respectively.

In noninferiority studies, the aim is to demonstrate that a device is noninferior to an active control with respect to a difference  $\delta$  deemed to be the smallest clinically meaningful difference. Unfortunately, a fundamental question that is sometimes not adequately addressed is whether noninferiority with respect to  $\delta$  implies that the device is effective, that is, superior to placebo. Because a placebo arm is frequently missing from active control trials but can be present in the trials that were the basis of approving the active control, the question of effectiveness can be addressed naturally by Bayesian methods (Simon, 1999). If the active control was not overwhelmingly superior to the placebo in the earlier trials, Bayesian methods could indicate considerable uncertainty as to whether the device is effective after accounting for variation between studies.

Bayesian approaches to multiplicity were considered for the specific applications of multiple subgroup testing (Section 5) and interim analysis (Section 6). Simultaneous testing for noninferiority and superiority relative to a control is another example of a fairly common multiple comparisons problem in device trials. For example, if the device could conceivably be superior to control in effectiveness, but would also be approvable if it were noninferior because of advantages in safety or reliability, then the company may want to test the two hypotheses simultaneously. Interestingly, because the two hypotheses are nested, simultaneous hypothesis testing does not inflate the overall Type I error rate relative to the individual test levels (Dunnett and Gent, 1996; Morikawa and

Yoshida, 1995). A Bayesian view is that any proper prior on the treatment effect  $\delta$  induces a more conservative decision rule for concluding superiority ( $\delta > 0$ ) than for concluding noninferiority ( $\delta > -\delta_0$ , for some  $\delta_0 > 0$ ). The reason is that the prior probability is always greater for noninferiority than for superiority, due to their nested structure (Pennello, 2003b).

Postmarket surveillance of medical products is an important mandate of FDA for which a Bayesian approach has been used for at least seven years. Surveillance of a device sometimes leads to regulatory action if the device is reported to have safety problems, as indicated by adverse event rates that are elevated relative to other devices of the same type. Unfortunately, surveillance is far from an exact science. Rates can be impossible to calculate because adverse events are often underreported and the denominator, number of devices that have been used, is not reported at all. Moreover, the possible number of combinations defined by the number of medical products that have been marketed and the types of adverse events that could be reported can be enormous, creating a difficult multiplicity problem with relatively scarce data. The Center for Drug Evaluation and Research (CDER) addresses the multiplicity problem in the surveillance of drugs with a Bayesian approach (DuMouchel, 1999). The denominator problem is handled by comparing relative rates among drugs in a class instead of the rates themselves. The relative rates are shrunk toward to the overall rate using a mixture model, which reduces greatly the number of signals that warrant further study. This approach may prove to be quite useful for automating the signaling of disproportional adverse event profiles for marketed medical devices.

## APPENDIX I – WINBUGS CODE FOR EXAMPLES

We give the WinBUGS code for examples, using WinBUGS format. In order to use an R interface (e.g., R2WinBUGS or BRugs), place the model code in a separate file, per the respective instructions for each program.

### 1) Cardiovascular Device Example

```

Model
{
  for (i in 1:S) {
    yy[i] ~ dbin(pp[i], nn[i]);
    logit(pp[i]) <- mu[i];
  }
  for (i in 1:S) {
    mu[i] ~ dnorm(mu0, tau0);
  }
  mu0 ~ dnorm(0, .001);
  tau0 ~ dgamma(.001, .001);
  # tau0 ~ dgamma(10,10);
}
#informative prior
#(hierarchical model)
#studies are exchangeable
#modification to discount prior
information

```

```

# for (i in 1:S) {
#           mu[i] ~ dnorm(0, .001);
# }
#noninformative prior
#studies are not exchangeable
#used to compute effective
#sample size

# mu0 ~ dnorm(0, .01);
# tau0 ~ dgamma(.01, .01);
for (i in 1:S) {
  diff[i] <- pp0 - pp[i];
  prob[i] <- step(diff[i]);
}
#diff between pp and OPC for study i

```

Data I. For Study 1 count is listed as missing (NA) to calculate prior probability of claim as prob[1].

```
list(nn = c(200, 135,25,1960,205,260,415),
yy = c(NA, 20, 5, 325, 43, 55, 60), S = 7, pp0 = .249)
```

Data II. For study 1, count is varied until success criterion is achieved (prob[1]>0.975), for purpose of calculating type I error rate and power.

```
list(nn = c(200, 135,25,1960,205,260,415),
yy = c(44, 20, 5, 325, 43, 55, 60), S = 7, pp0 = .249)
```

Initial starting values for parameters

```
list(mu0 = 1,tau0 = 1)
```

## 2) LIFE Study Example

Model

```

{
  for (i in 1:G) {
    vv[i] <- pow(se[i],2);
    tau[i] <- 1/vv[i];
    yy[i] ~ dnorm(mu[i], tau[i]);
    mu[i] ~ dnorm(mu0, tau.mu);
    prob[i] <- step(mu[i]);
    haz[i] <- exp(mu[i]);
  }
  mu0 ~ dnorm(0.0, eps);
  tau.mu ~dgamma(eps, eps);
}

```

#vv[i] = variance of hazard ratio yy[i]  
#tau[i] = precision of hazard ratio yy[i]  
#yy[i] = log hazard ratio for group i  
#exchangeability (random effects) modeling  
#prob[i] = probability that mu[i] > 0  
#haz[i] = hazard ratio  
#diffuse priors

Data from LIFE study

```
list(G = 2, yy = c(0.51019, -0.18730), se = c(0.23900, 0.06289), eps = .001)
```

Initial starting values for parameters

```
list(mu0 = 0, tau.mu = 5)
```

### 3) Model for Subgroup Analysis Confirmatory Study Example

Model

```
{
for (pp in 1:P) {
  yy[pp] ~ dnorm(mn[pp], tau.yy);
  mn[pp] <- omega[sdy[pp],trt[pp]] + xx[grp[pp],trt[pp]];
}
for (ss in 1:S) {
ksi[ss] ~ dnorm(0,tau.ksi)           #ksi = random study effects
for (tt in 1:T) {                   #tau = fixed treatment effects
  omega[ss,tt]~ dnorm(mn.omega[ss,tt],tau.st);
  mn.omega[ss,tt] <- ksi[ss] + tau[tt]; }} #tau.st = study by trt
                                           interaction variance

for (gg in 1:G) {
  gam[gg] ~ dnorm(0,tau.gam)         #gam = random group effects
for (tt in 1:T) {                   #tau.gt = grp by trt
  xx[gg,tt]~dnorm(gam[gg],tau.gt)}}  interaction variance

for (tt in 1:T) {tau[tt]~dnorm(0,0.0001)} #diffuse prior on fixed tau
                                           effects
tau.gam ~ dgamma(aa,bb);           #diffuse priors on
tau.ksi ~ dgamma(aa,bb);           #variance components
tau.gt ~ dgamma(aa,bb);
tau.st ~ dgamma(aa,bb);
tau.yy ~ dgamma(aa,bb);

for (ss in 1:S) {
for (gg in 1:G) {
for (tt in 1:T) {                   #mu = means
mu[ss,gg,tt] <- omega[ss,tt] + xx[gg,tt];}}}

for (ss in 1:S) {
for (gg in 1:G) {                   #mudif = mean differences
mudif[ss,gg] <- mu[ss,gg,2]-mu[ss,gg,1];}} #mudif[2,1] = parameter of
                                           interest
}
}
```

Inits

```
list(tau.yy = 1, tau.gam = 1, tau.ksi = 1, tau.gt = 1, tau.st = 1, tau = c(0,0))
```

Data

```
list(S = 2, G = 4, T = 2, P = 100, aa = 1, bb = 0.0025,
```

```

yy = c(
  0.28,0.29,0.21,0.19,0.31,0.21,0.28,0.26,0.27,0.24,0.21,0.24,0.22,0.13,0.25,
  0.21,0.27,0.10,0.23,0.15,
  0.25,0.19,0.21,0.20,0.16,0.12,0.20,0.26,0.19,0.23,0.24,0.19,0.24,0.29,0.16,
  0.30,0.21,0.27,0.20,0.16,
  0.16,0.25,0.30,0.22,0.20,0.24,0.20,0.16,0.32,0.30,0.17,0.17,0.24,0.17,0.31,
  0.27,0.29,0.10,0.30,0.25,
  0.15,0.19,0.16,0.25,0.17,0.22,0.22,0.19,0.21,0.15,0.24,0.20,0.26,0.21,0.19,
  0.28,0.25,0.22,0.29,0.20,
  0.30,0.25,0.26,0.19,0.21,0.26,0.23,0.21,0.34,0.30,0.24,0.23,0.26,0.19,0.29,
  0.17,0.21,0.15,0.13,0.25),
sdy = c(
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),
grp = c(
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
  4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
trt = c(
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2))

```

#### 4) Synthetic Control Example

```

Model
{
  for (i in 1:K) {
    yy[i] ~ dnorm(mu[i], tau);
    mu[i] <- alp[trt[i]] + diab[i]*bet;
  }
  tau ~ dgamma(.01, .01);
  bet ~ dnorm(0, .001);

  alp[1] ~ dnorm(0, .001) I(-10,100);
  for (t in 2:T) {
    alp[t] ~ dnorm(alp0, tau.alp); #historical controls exchangeable
  }
}

```



quantities as the posterior mean and variance above. An  $\alpha$ -level test of  $H$  based on  $\hat{\mu}$  is to reject  $H$  if  $(c + d)^{1/2}(\hat{\mu} - \mu_0) > z_{1-\alpha}$ , i.e.,  $(c + d)^{-1/2}(c^{1/2}z_x + d^{1/2}z_y) > z_{1-\alpha}$ . That is, the frequentist  $\alpha$ -level test and the Bayesian decision rule are the same.

### Type I Error: View 2

When considering the study alone, an  $\alpha$ -level test of  $H$  is to reject  $H$  if  $z_y > z_{1-\alpha}$ . Now suppose we have very strong prior information, say,  $z_x > kz_{1-\alpha}$ , where  $k$  is large. Then by requiring  $z_y > z_{1-\alpha}$  to reject  $H$ , the test statistic  $(c + d)^{-1/2}(c^{1/2}z_x + d^{1/2}z_y)$  used when prior information  $x$  and data  $y$  are considered together has to be much larger than  $z_{1-\alpha}$  to reject  $H$ . In fact, if  $c = d$ , it has to be larger than  $2^{-1/2}(k + 1)z_{1-\alpha}$ . From the Bayesian perspective, to reject  $H$  the posterior probability of  $H$  has to be less than  $\Phi(2^{-1/2}(k + 1)z_{1-\alpha})$ , which for  $\alpha = 0.05$  (one-sided) is 0.9900, 0.9998, and 1.0000 for  $k = 1, 2, 3$  respectively.

### ACKNOWLEDGMENT

No official support or endorsement by the Food and Drug Administration of this paper is intended or should be inferred.

### REFERENCES

- Albert, P., Dodd, L. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 60:427–435.
- Beiden, S. V., Campbell, G., Meier, K. L., Wagner, R. F. (2000a). On the problem of ROC analysis without truth : the EM algorithm and the information matrix. *Proc. SPIE* 3981:126–134.
- Beiden, S. V., Wagner, R. F., Campbell, G. (2000b). Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects, receiver operating characteristic analysis. *Acad. Radiol.* 7:341–349.
- Begg, C. (1990). On inferences from Wei's biased coin design for clinical trials. *Biometrika* 77:467–484.
- Berry, D. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Stat. Sci.* 19:175–187.
- Black, M. A., Craig, B. A. (2002). Estimating disease prevalence in the absence of a gold standard. *Stat. Med.* 21:2653–2669.
- Broemeling, L. (2007). *Bayesian Biostatistics and Diagnostic Medicine*. Boca Raton, FL: CRC Press.
- Carlin, B. P., Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.
- Dendukuri, N., Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 57:158–167.
- Dixon, D. O., Simon, R. (1991). Bayesian subset analysis. *Biometrics* 47:871–881.
- Dixon, D. O., Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial. *Stat. Med.* 11:13–22.
- Dunnett, C. W., Gent, M. (1996). An alternative to the use of two-sided tests in clinical tests. *Stat. Med.* 15:1927–1738.
- DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* 53:177–202.

- Gelfand, A. E., Sahu, S. K., Carlin, B. P. (1995). Efficient parameterizations for normal linear mixed models. *Biometrika* 82:479–488.
- Gelman, A. (2006). Prior distributions for variance components in hierarchical models (Comment on article by Browne and Draper). *Bayesian Stat.* 1:515–534.
- Gelman, A., Meng, X., Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* 6:733–760 (discussion: 760–807).
- Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman & Hall.
- Georgiadis, M. P., Johnson, W. O., Singh, R., Gardner, I. A. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl. Stat.* 52:63–76.
- Greenes, R. A., Begg, C. B. (1985). Assessment of diagnostic technologies: Methodologies for unbiased estimation from samples of selectively verified patients. *Invest. Radiol.* 20:751–756.
- Hoppin, J. W., Kupinski, M. A., Kastis, G. A., Clarkson, E., Barrett, H. H. (2002). Objective comparison of quantitative imaging modalities without the use of a gold standard. *IEEE T. Med. Imaging* 21:441–449.
- Ibrahim, J. G., Chen, M. H. (2000). Power prior distributions for regression models. *Stat. Sci.* 15:46–60.
- Irony, T. Z., Pennello, G. A. (2001). Choosing an appropriate prior for Bayesian medical device trials in the regulatory setting. 2001 Proceedings of the Biopharmaceutical Section [CD-ROM], Alexandria, Virginia: American Statistical Association, Vol. M, 85.
- Jennison, C., Turnbull, B. (2000). *Group Sequential Methods with Application to Clinical Trials*. London: Chapman & Hall.
- Johnson, T. D., Johnson, V. E. (2006). A Bayesian hierarchical approach to multirater correlated ROC analysis. *Stat. Med.* 25:1858–1871.
- Kosinski, A., Barnhart, H. (2003). A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Stat. Med.* 22:2711–2721.
- Kupinski, M. A., Hoppin, J. W., Clarkson, E., Barrett, H. H., Kastis, G. A. (2002). Estimation in medical imaging without a gold standard. *Acad. Radiol.* 9:290–297.
- Malec, D. (2001). A closer look at combining data among a small number of binomial experiments. *Stat. Med.* 20:1811–1824.
- Morikawa, T., Yoshida, M. (1995). A useful testing strategy in Phase III trials: combined test of superiority and test of equivalence. *J. Biopharm. Stat.* 5:297–306.
- Ohlssen, D., Sharples, L., Spiegelhalter, D. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Stat. Med.* 26:2088–2112.
- O'Malley, J., Normand, S., Kuntz, R. (2002). Sample size calculation for a historically controlled clinical trial with adjustment for covariates. *J. Biopharm. Stat.* 12:227–247.
- Pennello, G. A. (2003a). Comparing monitoring devices when a gold standard is unavailable: application to pulse oximeters. 2003 Proceedings of the American Statistical Association, Biopharmaceutical Section [CD-ROM], Alexandria, VA: American Statistical Association, pp. 3256–3263.
- Pennello, G. (2003b). Comments and rejoinder on “Issues of simultaneous tests for noninferiority and superiority”. *J. Biopharm. Stat.* 13:641–662.
- Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: A short review. *Stat. Meth. Med. Res.* 12:489–504.
- Pocock, S. (1976). The combination of randomized and historical controls in clinical trials. *J. Chron. Dis.* 29:175–188.
- Rosenbaum, P., Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

- Scott, P. E., Campbell, G. (1998). Interpretation of subgroup analyses in medical device clinical trials. *Drug Inf. J.* 32:213–220.
- Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics* 55:484–487.
- Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Stat. Soc. B.* 64:583–640.
- Wang, F., Gelfand, A. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Stat. Sci.* 17:193–208.