Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

SAMPLE SIZE CALCULATION FOR A HISTORICALLY CONTROLLED CLINICAL TRIAL WITH ADJUSTMENT FOR COVARIATES

A. James O'Malley [a] , Sharon-Lise T. Normand [b] & Richard E. Kuntz [c]

[a] Department of Health Care Policy , Harvard Medical School , 180 Longwood Avenue, Boston, MA, 02115, U.S.A.

[b] Department of Health Care Policy , Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health , Boston, Massachusetts, USA, U.S.A.

[c] Divisions of Clinical Biometrics and Cardiovascular Disease, Department of Medicine , Brigham and Women's Hospital, Harvard Medical School , Boston, Massachusetts, USA, U.S.A.

Published online: 05 Oct 2011.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 12, No. 2, pp. 227–247, 2002

SAMPLE SIZE CALCULATION FOR A HISTORICALLY CONTROLLED CLINICAL TRIAL WITH ADJUSTMENT FOR COVARIATES

A. James O'Malley,[1,*] Sharon-Lise T. Normand,[2] and Richard E. Kuntz[3]

[1]Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA
[2]Department of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA
[3]Divisions of Clinical Biometrics and Cardiovascular Disease, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

ABSTRACT

We present a Bayesian approach to determining the optimal sample size for a historically controlled clinical trial. This work is motivated by a trial of a new coronary stent that uses a retrospective control group formed from seven trials of coronary stents currently marketed in the United States. In studies involving nonrandomized control groups, hierarchical regression, propensity score methods, or other sophisticated models are typically required to account for heterogeneity among groups which, if ignored could bias the results. Sample size calculations for historically controlled trials of medical devices are often based on formulae derived for randomized trials and fail to account for estimation of model parameters, correlation of observations, and uncertainty in the distribution of covariates of the patients recruited in the new trial. We propose methodology based

*Corresponding author. Fax: (617) 432-0173; E-mail: omalley@hcp.med.harvard.edu

DOI: 10.1081/BIP-120015745
1054-3406 (Print); 1520-5711 (Online)

on stochastic optimization that overcomes these deficiencies. The methodology is demonstrated using an objective function based on the power of the trial from a Bayesian approach. Analytic approximations based on a covariate-free analysis that convey features of the power function are developed. Our principle conclusions are that exact sample size calculations can be substantially different from current approximations, and stochastic optimization provides a convenient method of computation.

Key Words: Cardiovascular disease; Stents; Bayesian hierarchical modeling; Sample size; Covariates

INTRODUCTION

In this article, we develop sample-size methodology when the historical data are grouped, possibly arising from multiple trials, and when several observed covariates are known to affect the outcome. To retain familiarity with standard practices, we use the power of the trial as the basis of all calculations.

Historically controlled trials are studies in which investigators compare, retrospectively, current patients on a new treatment with previous or "historical" patients on the standard treatment. Because such studies are observational in design, it is important that known differences between treatment groups are controlled for using regression adjustment or matching methods. This necessarily complicates the analysis of the trial. If complexities of the intended analysis are not accounted for in the design of the trial, the study may be severely underpowered, increasing the likelihood that the recruitment of insufficient patients leads to a negative result, or over-powered; unnecessarily wasting financial resources, patient involvement, and time.

Previous research has employed analytic approximations to derive power and sample-size formulae for regression models (for example, Ref. [1] for generalized linear models; Ref. [2,3] for logistic regression; Ref. [4] for Poison regression). Because the derivation of accurate analytic approximations does not seem practical, numerical methods offer a sensible solution. We adapt the algorithm for stochastic optimization in Ref. [5] to compute power probabilities. The incorporation of uncertainty in parameter values, inclusion of covariate effects, and the use of precise numerical methods rather than analytic approximations, distinguish this from previous work.

In the remainder of this section, we describe the clinical trial of a new coronary-artery stent, which motivated our work. Although the trial involves a binary outcome and a study protocol in which conclusions are based on hypothesis tests, the general methodology proposed adapts easily to other types of outcome variables and modes of inference. In the section "General Framework" we develop the methodology for power and optimal sample size calculations using the

coronary stent problem for illustration. In the section "Example: Coronary-Artery Stents" computations for and results related to the coronary stent data are presented. Concluding remarks are given in the section "Discussion".

Coronary-Artery Stent Trials and Designs

A coronary stent is a thin expandable metallic tube that is inserted into a coronary artery in order to widen an obstructed coronary segment. Stent modifications are introduced to improve upon operative-procedure features or costs, and impact only minimally on measurable patient outcomes. This is in contrast to pharmaceuticals where a small alteration in the molecule may result in profound changes in patient outcomes. As a result, to reduce the length of the trial, researchers have used single arm studies in which all patients receive the new stent and comparisons are made to a historical control group.[6]

We illustrate our methodology using a trial of a new stent for which a historical control group is formed from seven randomized trials of stents currently marketed in the United States. The historical data (Table 1 and in the section "Analyzing Coronary Stent Data") include seven covariates and a binary end-point measuring restenosis. A similar data set has been analyzed previously by O'Malley et al.[7]

Suppose there are n historical trials. In an $(n+1)$th trial of a new stent, the key test of interest is the "noninferiority" test[8]

$$\mathcal{H}_0\colon \quad p_{n+1} \geq p_{\mathrm{h}} + \omega_0 \quad \text{vs.} \quad \mathcal{H}_{\mathrm{A}}\colon \quad p_{n+1} < p_{\mathrm{h}} + \omega_0,$$

where $\omega_0 \geq 0$ represents the minimum clinically significant difference, and p_{h} and p_{n+1} denote the failure probabilities associated with the historical and new devices. If $\mathcal{H}_{\mathrm{A}}$ is concluded, the device is approved for release in the US market with a label stating that the new device is "safe and effective compared to industry standards"[9,10].

Current methods of determining sample sizes for historically controlled trials of medical devices are based on a simple formula derived from Refs. [8,11] Suppose that m_{h} Bernoulli outcomes yielding a proportion $\hat{p}_{\mathrm{h}}$ of failures are available from the historical trials, denote the type I error of the test by α, and let $p_{n+1} - p_{\mathrm{h}} = \omega_{\mathrm{A}}$ be the assumed treatment effect. The power at sample size m_{n+1} is estimated using

$$\mathrm{pow}(m_{n+1}, \omega_{\mathrm{A}}, p_{\mathrm{h}} | \hat{p}_{\mathrm{h}}) =$$

$$\Phi\left[\frac{\hat{p}_{\mathrm{h}} - p_{\mathrm{h}} + \omega_0 - \omega_{\mathrm{A}} - z_\alpha \{p_{\mathrm{h}}(1-p_{\mathrm{h}})/m_{\mathrm{h}} + (p_{\mathrm{h}}+\omega_0)(1-p_{\mathrm{h}}-\omega_0)/m_{n+1}\}^{1/2}}{\{(p_{\mathrm{h}}+\omega_{\mathrm{A}})(1-p_{\mathrm{h}}-\omega_{\mathrm{A}})/m_{n+1}\}^{1/2}}\right], \tag{1}$$

where z_α is the $100(1-\alpha)$ percentile of the standard normal distribution.

Table 1. Historical Data for Patients Participating in Seven Randomized Coronary-Artery Stent Trials Conducted Between 1993 and 1998 Involving 5806 Patients

Trial	No. of Patients	TLR (Proportion)	Length (mm)	Diameter (mm)	Thickness (Ratio)	Diabetes (%)	Location (%)	No. of Diseased Vessels	
								2 (%)	3 (%)
1	1806	0.116	11.57 (6.27)	2.99 (0.49)	0.06 (0.13)	18	43	26	11
2	1248	0.083	11.36 (6.05)	3.03 (0.48)	0.08 (0.11)	20	42	22	7
3	201	0.1	14.72 (7.90)	3.02 (0.47)	0.09 (0.10)	21	40	24	11
4	1087	0.107	13.97 (7.52)	3.03 (0.53)	0.08 (0.12)	24	40	28	13
5	925	0.114	13.91 (8.47)	3.03 (0.49)	0.07 (0.13)	19	43	24	10
6	313	0.099	12.74(5.84)	3.11 (0.48)	0.1 (0.10)	19	38	23	11
7	226	0.097	14.06 (7.08)	3.06 (0.59)	0.03 (0.13)	23	39	21	13
Overall	5806	0.104	12.62 (7.04)	3.02 (0.50)	0.07 (0.12)	20	42	25	10

Means (standard deviations) are reported for the continuous variables; percentages are reported for the binary variables.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

The computation performed in Eq. (1) is routinely applied even if adjustments for covariates are planned. With data derived from multiple trials, both the analysis and hence sample size calculation should also account for the correlation between outcomes from the same trial.

In describing the distribution of data from the new trial, it is not necessary that p_h be specified. In practice, it seems prudent to account for the uncertainty in p_h rather than conditioning on a point estimate such as $\hat{p}_h$.[12,13] This allows the uncertainty in the parameters to propagate through the design calculations, and avoids making arbitrary assumptions about the effect size. We adopt a Bayesian approach and evaluate the joint probability of any possible conclusion and its correctness using a probability distribution on the parameter of interest. That is, we integrate Eq. (1) with respect to the posterior distribution of p_h given the historical data. Point specification is a special case of this methodology.

GENERAL FRAMEWORK

Analyzing Coronary Stent Data

Table 1 summarizes the data to be used to design a historically controlled trial of a new stent. The outcome and predictor variables are briefly described (see Ref. [7] for further details). The primary endpoint, target lesion revascularization (TLR), is adjudicated by a clinical events committee who is blinded to treatment assignment. TLR indicates whether the patient had to undergo repeat angiography within 270 days of the stent implant so that large values of TLR indicate poor stent performance.

Six important predictors of TLR are displayed in Table 1: lesion length in millimeters (length); diameter of healthy reference vessel in millimeters (diameter); plaque thickness reported as a fraction of diameter immediately after stent delivery (thickness); history of diabetes (diabetes); presence of disease in the left anterior descending coronary artery (location); and the number of diseased vessels (1, 2, or 3). The number of diseased vessels is represented in regression models as two binary variables using one vessel as the reference category. The mean values of the predictor variables and the prevalence of TLR varies among trials. Presence of each binary risk factor, length, and thickness are positively correlated with TLR while diameter is negatively correlated.

The subscripts i and j ($i = 1, \ldots, n; j = 1, \ldots, m_i$) index the trial and patient within trial respectively, where m_i is the total number of patients in the ith trial, and n is the number of historical trials. Let $y_{ij} = 1$ if TLR occurred for the jth patient in the ith trial and 0 otherwise, and x_{ij} be a $p \times 1$ vector of predictor variables centered about their overall mean so that $\sum_{i,j} x_{ij} = 0$. Let $f(\cdot)$ denote a probability distribution of the corresponding random variable. The following

model describes outcomes in the historical trials:

$$y_{ij}|x_{ij}, \theta_i, \beta \overset{\text{ind}}{\sim} \text{Bernoulli}(y_{ij}|\eta_{ij})$$

where

$$\text{logit}(\eta_{ij}) = \log\left(\frac{\eta_{ij}}{1-\eta_{ij}}\right) = \theta_i + \beta^T x_{ij} \quad \text{for all } i, j, \tag{2}$$

$$\theta_i|\mu, \tau^2 \overset{\text{ind}}{\sim} \text{N}(\theta_i|\mu, \tau^2) \quad \text{for all } i, \beta \sim \text{N}(\beta|b_0, S_0), \tag{3}$$

$$\mu \sim \text{N}(\mu|a_0, s_0^2), \tau^2 \sim \text{IG}(\tau^2|u_0, v_0), \tag{4}$$

where N(·) and IG(·) denote Gaussian and Inverse Gamma densities, respectively; θ_i is a random trial-level main effect drawn from a hypothetical population with mean μ and variance τ^2; and β is a $p \times 1$ vector of regression parameters. Hyperparameters specified by the investigator are denoted by roman letters, subscripted by 0.

The hierarchical generalized linear model defined in Eqs. (2)–(4) is referred to as a *random-intercept* model.[14] Information about β is obtained only from comparisons within trials because the θ_i-term absorbs all between-trial information. Preliminary investigations[7] indicated that allowing β to vary among trials failed to improve the model fit.

The outcomes for a new $(n+1)$th trial of m_{n+1} patients are described by

$$y_{(n+1)j}|x_{(n+1)j}, \theta_{n+1}, \beta \overset{\text{ind}}{\sim} \text{Bernoulli}(y_{ij}|\eta_{(n+1)j})$$

where

$$\text{logit}(\eta_{(n+1)j}) = \theta_{n+1} + \beta^T x_{(n+1)j} \quad \text{for all } j, \tag{5}$$

$$\theta_{n+1} \sim \text{N}\left(\theta_{n+1}|m_0, t_0^2\right). \tag{6}$$

Because we ultimately want to compare the outcomes from the new trial with the outcomes from the historical trials and do not want to favor either hypothesis in the model specification, θ_{n+1} and $\{\theta_i\}_{1:n}$ are a priori independent. Therefore, the only learning about θ_{n+1} from the historical trials is via the regression coefficient β.

In a generalized linear models framework, it is natural for hypotheses to be constructed on the scale of the linear predictor. The noninferiority test

$$\mathcal{H}_0: \theta_{n+1} \geq \bar{\theta}_{\text{h}} + \delta_0 \quad \text{vs.} \quad \mathcal{H}_{\text{A}}: \theta_{n+1} < \bar{\theta}_{\text{h}} + \delta_0, \tag{7}$$

where $\bar{\theta}_{\text{h}} = \sum_{i=1}^{n} \theta_i/n$, and $\delta_0 > 0$ is the minimum clinically significant difference translated from ω_0 in Eq. (1) to the log–odds scale, is assumed. At a fixed value of the covariates, the log–odds ratio of the outcomes in the new and historical trials reduces to $\theta_{n+1} - \bar{\theta}_{\text{h}}$. Hence, the test in Eq. (7) applies to all patients. Although

Eq. (7) does not involve (β, μ, τ^2), these parameters affect the outcome and therefore impact the power of the test.

An alternative test is obtained by substituting $\bar{\theta}_h$ with μ in Eq. (7). This is appropriate when the historical trials represent a sample from a population of the same type of trials.

Decision Rule

We use the test defined in Eq. (7) with the decision rule: conclude $\mathscr{H}_A$ if the posterior probability of $\mathscr{H}_A$ exceeds $1 - \alpha$. Let D_h denote all historical data, $D_h = \{y_{ij}, x_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i\}$. The new trial data are denoted by $y_{n+1} = (y_{(n+1)1}, \ldots, y_{(n+1)m_{n+1}})$ and $X_{n+1} = (x_{(n+1)1}, \ldots, x_{(n+1)m_{n+1}})$. The trial main effects are denoted by $\theta = (\theta_h, \theta_{n+1})$, where $\theta_h = (\theta_1, \ldots, \theta_n)$ is the vector of random effects for the historical trials, and remaining parameters by $\xi = (\beta, \mu, \tau^2)$. Further, let $\Theta_A = \{\theta : \theta_{n+1} < \bar{\theta}_h + \delta_0\}$ be the set of values of θ for which $\mathscr{H}_A$ holds. Then $\mathscr{H}_A$ is concluded if

$$\Pr(\mathscr{H}_A | y_{n+1}, X_{n+1}, D_h) = \frac{\int_{0 \in \Theta_A, \xi} f(y_{n+1}|X_{n+1}, \theta_{n+1}, \xi) f(\theta_{n+1}) f(\theta_h, \xi|D_h) \mathrm{d}(\xi, \theta)}{f(y_{n+1}|X_{n+1}, D_h)}$$

$$\geq 1 - \alpha,$$

where $f(y_{n+1}|X_{n+1}, D_h) = \int_{0 \in \xi} f(y_{n+1}|X_{n+1}, \theta_{n+1}, \xi) f(\theta_{n+1}) f(\theta_h, \xi|D_h) \mathrm{d}(\xi, \theta)$ is the marginal distribution of $(y_{n+1}|X_{n+1}, D_h)$.

The set of values of y_{n+1} for which $\mathscr{H}_A$ is concluded, given (X_{n+1}, D_h), is denoted by

$$\mathscr{R}_A(X_{n+1}, D_h) = \{y_{n+1} : \Pr(\mathscr{H}_A | y_{n+1}, X_{n+1}, D_h) \geq 1 - \alpha\}.$$

The space identified by $\mathscr{R}_A = \mathscr{R}_A(X_{n+1}, D_h)$ is often referred to as the rejection region.

Determining Power

The power of the trial at (X_{n+1}, θ, ξ) is the integral of $f(y_{n+1}|X_{n+1}, \theta, \xi)$, the likelihood function for the outcomes in the new trial, over $\mathscr{R}_A$. This is given by:

$$\mathrm{pow}(X_{n+1}, \theta, \xi|D_h) = \int_{y_{n+1} \in \mathscr{R}_A(X_{n+1}, D_h)} f(y_{n+1}|X_{n+1}, \theta, \xi) \mathrm{d}y_{n+1}. \quad (8)$$

The expression obtained in Eq. (8) is a function of unknown parameters (θ, ξ) and future covariates X_{n+1}. In a traditional power calculation, specific values would be substituted for (θ, ξ), and for terms involving X_{n+1}. Our approach is more general in that the uncertainty in these terms is incorporated.

Let $w(X_{n+1}, \theta, \xi|D_h) = w(X_{n+1}|D_h)w(\theta, \xi|X_{n+1}, D_h)$ be a probability distribution function, representing the importance of different values of (X_{n+1},θ,ξ). In general the specification of $w(\cdot)$ may be relaxed to that of a utility function, however, this is not necessary for this application. To avoid ambiguity with the probability distributions involved in the analysis of the trial, we refer to $w(\cdot)$ as a parameter weighting distribution.

Uncertainty in (θ,ξ) is incorporated by integrating the product of $\text{pow}(X_{n+1}, \theta, \xi|D_h)$ and $w(\theta,\xi|X_{n+1},D_h)$ over $\theta \in \Theta_A$ and ξ, yielding

$$\text{pow}(X_{n+1}|D_h) = \int_{\theta\subset\Theta_{A,\xi}} \text{pow}(X_{n+1}, \theta, \xi|D_h)w(\theta, \xi|X_{n+1}, D_h)\text{d}(\theta, \xi). \quad (9)$$

Equation (9) computes power for fixed values of the covariate vectors in X_{n+1}, corresponding to hypothetical patients in the new trial. To incorporate uncertainty in X_{n+1}, $\text{pow}(X_{n+1}|D_h)$ is averaged over $w(X_{n+1}|D_h)$ to give the power of the trial for sample size m_{n+1}:

$$\text{pow}(m_{n+1}|D_h) = \int_{X_{n+1}} \text{pow}(X_{n+1}|D_h)w(X_{n+1}|D_h)\text{d}X_{n+1}. \quad (10)$$

Choice of Parameter Weighting Distribution

In the trial of a new stent, θ_{n+1} (specifically $\theta_{n+1} - \bar{\theta}_h$), is of primary interest while (X_{n+1}, θ_h, ξ) may be viewed as a set of nuisance terms. We weight (θ_h,ξ) by $f(\theta_h, \xi|D_h)$, the prior based on the historical data D_h, and take $w(X_{n+1}|D_h)$ to be the probability distribution of the covariates for the new trial, denoted by $f(X_{n+1}|D_h)$. This assumes that (θ_h, ξ) and X_{n+1} are independent given D_h, an assumption that seems reasonable. This parameter weighting distribution has the form

$$w(X_{n+1}, \theta, \xi|D_h) = f(X_{n+1}|D_h)f(\theta_h, \xi|D_h)w(\theta_{n+1}|X_{n+1}, \theta_h, \xi, D_h).$$

The covariate vectors in D_h may be used to determine $f(X_{n+1}|D_h)$. A simple, but often realistic choice, is defined by

$$\Pr(x_{(n+1)k} = x_{ij}|D_h) = \frac{1}{N}(i = 1, \ldots, n;\ j = 1, \ldots, m_i;\ k = 1, \ldots, m_{n+1}),$$

where

$$N = \sum_{i=1}^{n} m_i.$$

In determining the sample size for a new trial, a common practice is to evaluate power at specific values of an effect size of interest, in this case $\theta_{n+1} - \bar{\theta}_h$. Let $\Theta_{n+1} \subseteq \Theta_A$ denote the set of values of θ for which $\theta_{n+1} - \bar{\theta}_h \in \delta_A(\theta)$. The set $\delta_A(\theta)$ can range from a single point δ_A (such that $\delta_A < \delta_0$), as for a traditional

power calculation, to $(-\infty,\delta_0)$. The latter case yields the predictive power of Spiegelhalter and Freedman[12] averaged over the uncertainty in ξ and X_{n+1}.

The parameter weighting distribution for θ_{n+1} is denoted $w(\theta_{n+1}|\theta_h)$ as it depends only on θ_h. This assigns positive weights to $\theta \in \Theta_{n+1}$ and zero weight to $\theta \notin \Theta_{n+1}$. In "Comparative Bayesian Computations" we assume that

$$w(\theta_{n+1}|\theta_h) = I(\theta \in \Theta_{n+1}), \tag{11}$$

where $I(\cdot)$ is the indicator function, and consider point and interval specifications of $\delta_A(\theta)$ (see the section "Comparative Bayesian Computations").

Implementation

Numerical integration is required to evaluate the expressions in Eqs. (9) and (10). We follow Müller and Parmigiani[5] in applying stochastic optimization via curve fitting Monte Carlo samples to estimate the optimal sample size. The algorithm proceeds by:

1. For simulation $s = 1, \ldots, S$:
   (a) Select a positive integer m^s_{n+1} (determined from a designed experiment).
   (b) Draw (i) X^s_{n+1} from $f(X_{n+1}|D_h)$, and (ii) (θ^s_h, ξ^s) from $f(\theta_h, \xi|D_h)$; followed by (iii) θ^s_{n+1} from $w(\theta_{n+1}|\theta^s_h)$; followed by (iv) y^s_{n+1} from $p(y_{n+1}|X^s_{n+1}, \theta^s_{n+1}, \xi^s)$.
   (c) Compute (i) $p^s = \Pr(\mathcal{H}_A|y^s_{n+1}, X^s_{n+1}, D_h)$, followed by (ii) $u^s = I(p^s \geq 1 - \alpha)$.
2. Fit a smooth one-dimensional curve $u(m_{n+1})$ to the simulated data $\{u^s, m^s_{n+1}\}_{1:S}$.
3. Evaluate deterministically $\hat{m}_{n+1}$, the smallest value of m_{n+1} for which $\text{pow}(X_{n+1}, D_h|\Theta_{n+1}) \geq 1 - \gamma$.

By the strong law of large numbers the long-run Monte Carlo average of u^s converges to the power at m^s_{n+1}.

Variability of the fitted power function and hence the uncertainty about $\hat{m}_{n+1}$ depends on the length of the simulation. A sufficiently long simulation allows the optimal sample size, the power at a certain sample size, or the power function itself to be estimated within a required level of precision (Ref. [5] Proposition 1). The nonparametric bootstrap[15] can be used to estimate the standard error of estimates based on simulation data.

Steps 1(b)(ii) and 1(c)(i) of the algorithm involve sampling from analytically intractable posterior distributions, and evaluation of high-dimensional integrals. Numerical methods are required for both steps. Samples from $f(\theta_h, \xi|D_h)$ are drawn by applying Markov Chain Monte Carlo (MCMC) methods to fit the model in Eqs. (2)–(4) corresponding to an analysis of the historical trials alone. MCMC methods are also used to fit the model for the analysis of a new trial, i.e.,

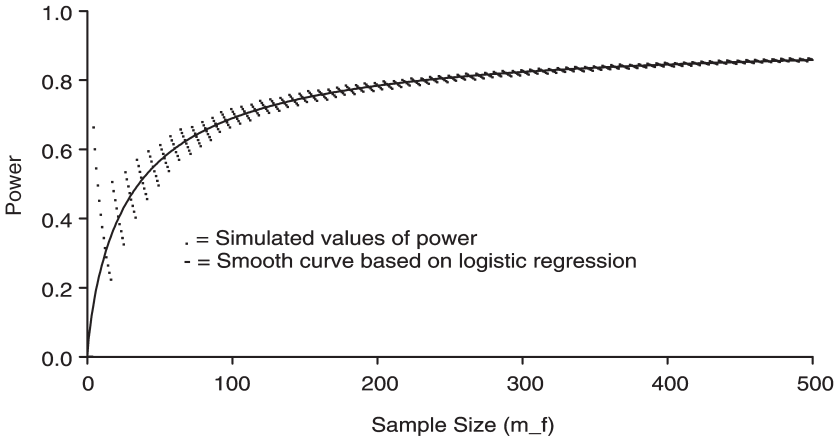

Figure 1. Power curve for the Bayesian Nonhierarchical Covariate-Free Model assuming uniform parameter weighting distributions defined on: $\Theta_{n+1} = \{\theta : -0.87 \leq \theta_{n+1} - \bar{\theta}_{\mathrm{h}} \leq 0.63\}$ (w_4). The historical control comprised 5806 patients with $\hat{p}_{\mathrm{h}} = 0.104$. The simulated values of power for $m_{n+1} \in \{0, 1, \ldots, 500\}$, denoted by dots are displayed along with a smooth curve based on the logistic regression model.

Eqs. (2)–(6), to each combined data set $(y^s_{n+1}, X^s_{n+1}, D_{\mathrm{h}})$. Monte Carlo averaging is employed to compute $\Pr(\mathcal{H}_{\mathrm{A}}|y^s_{n+1}, X^s_{n+1}, D_{\mathrm{h}})$ (Appendix A).

The outcome of the sth simulated analysis, u^s, is a Bernoulli random variable with expectation depending on the probability $\Pr(u^s = 1|m^s_f) = \mathrm{pow}(m^s_{n+1}, D_{\mathrm{h}}|\Theta_{n+1})$. An analytic expression for the power function can be derived by fitting a regression model to $\{u^s, m^s_{n+1})_{1:S}$. The logistic regression model with

$$\Pr(u^s = 1|m^s_f) = \frac{\exp\{\nu_0 + \nu_1 m^s_{n+1} + \nu_2 \log(m^s_{n+1})\}}{1 + \exp\{\nu_0 + \nu_1 m^s_{n+1} + \nu_2 \log(m^s_{n+1})\}}, \quad s = 1, 2, \ldots, S, \tag{12}$$

provides an accurate monotone-increasing estimate of the true power function for coronary stent trials (see the section "True Power Function" and Fig. 1). The parameter estimates of ν are substituted into Eq. (12) to evaluate power at different sample sizes. The smallest integer, $\hat{m}_{n+1}$, for which power is at least $1 - \gamma$ is found by solving $\Pr(u = 1|m_f) = 1 - \gamma$ for m_{n+1} and rounding upwards to the nearest integer. The solution typically needs to be obtained numerically.

When covariates are ignored, at fixed values of θ the power function can either be increasing or decreasing in m_{n+1}, and converge to 1 or 0, respectively, as $m_{n+1} \to \infty$ (Appendix B, Result 1). When the uncertainty in θ is accounted, the power function converges to a probability strictly between 0 and 1 (Appendix B, Result 2).

EXAMPLE: CORONARY-ARTERY STENTS

We evaluated Eq. (1) by setting $p_h = \hat{p}_h = 0.104$, the overall TLR rate for the historical trials in Table 1; and $\omega_0 = 0.075$, a minimum clinically significant difference agreed to by manufacturers and the FDA. The resulting power function at a type I error probability of 0.05 is

$$\text{pow}(m_{n+1}, \omega_A, p_h = 0.104 | \hat{p}_h = 0.104) = \Phi\left[\frac{0.075 - z_{0.05}(0.0932/5806 + 0.1470/m_{n+1})^{1/2}}{(0.0932/m_{n+1})^{1/2}}\right]. \quad (13)$$

When $\omega_A = 0$, i.e., the new stent is equivalent to the historical stents, the smallest sample size for which power exceeds 0.80 is 139 patients. When $\omega_A = 0.033$, i.e., the new stent is expected to perform slightly worse than the historical stents, a sample size of $m_{n+1} = 495$ is required to achieve a power of 0.80.

Comparative Bayesian Computations

At a specific value of p_h, $\omega = p_{n+1} - p_h$ corresponds to an effect size of

$$\delta = \log\{(p_h + \omega)/(1 - p_h - \omega)\} - \log\{p_h/(1 - p_h)\}, \quad (14)$$

on the log–odds scale. Therefore, the test

$$\mathcal{H}_0 : \theta_{n+1} \geq \bar{\theta}_h + 0.63 \quad \text{vs.} \quad \mathcal{H}_A : \theta_{n+1} < \bar{\theta}_h + 0.63$$

is comparable to that in Eq. (13) when $\omega_0 = 0.075$ and $p_h = 0.104$. Similarly, $\delta = 0.315$ corresponds to $\omega = 0.033$.

Three different Bayesian models are considered.

1. *Bayesian nonhierarchical covariate-free model*: $\theta_i = \theta_*\,(i = 1, \ldots, n)$ and $\beta = 0$. This model is the Bayesian analog of the model proposed in Eq. (1). It assumes that (a) the TLR rates of subjects are the same across the historical trials; and (b) no covariates affect TLR.
2. *Bayesian hierarchical covariate-free model*: θ unrestricted and $\beta = 0$. This model assumes (a) within each trial, the TLR rates are conditionally independent among subjects and depend on an unobserved trial-specific effect; and (b) the outcomes do not depend on subject level covariates.
3. *Bayesian hierarchical covariate-adjusted model*: θ and β unrestricted. This model assumes (a) within each trial, the risk of TLR is affected by subject characteristics and an unobserved trial-specific effect; and (b) conditional on the trial level effects and subject level covariates, TLR rates are independent.

In subsequent analyses, we take the empirical distribution of D_h given in Eq. (11) as $f(X_{n+1}|D_h)$, use $f(\theta_h, \xi|D_h)$ as the parameter weighting distribution for (θ_h, ξ), and assume that $w(\theta_{n+1}|\theta_h)$ is the uniform distribution defined on the set Θ_{n+1} as in Eq. (11). Four specifications of $\delta_A(\theta)$ are used:

$w_1 : \delta_A(\theta) = \{\theta_{n+1} : \theta_{n+1} - \bar{\theta}_h = 0\}$: the new stent is equivalent to the control stents.

$w_2 : \delta_A(\theta) = \{\theta_{n+1} : \theta_{n+1} - \bar{\theta}_h = 0.315\}$: the new stent is inferior to the control stents but not by an amount that is clinically significant.

$w_3 : \delta_A(\theta) = \{\theta_{n+1} : -1.50 \leq \theta_{n+1} - \bar{\theta}_h < 0\}$: the new stent is better than the control stents.

$w_4 : \delta_A(\theta) = \{\theta_{n+1} : -0.87 \leq \theta_{n+1} - \bar{\theta}_h < 0.63\}$: the new stent could be better or worse than the control stents.

Implementation of Bayesian Models

In the analysis of each simulated data set, independent $N(0, 10^6)$ prior distributions are assumed for $(\theta_{n+1}, \{\beta_i\}_{1:p}, \mu)$ (p is the size of β), and an independent $IG(10^{-3}, 10^{-3})$ prior distribution is assumed for τ^2. Noninformative priors are assumed for two reasons: they are easier to justify to regulatory authorities, and they provide a fairer comparison with the power calculation currently used for stent trials. It is important to note that the $N(0, 10^6)$ prior for θ_{n+1} is not used to generate values of θ_{n+1} for the new trial. This is the role of the parameter weighting distribution $w(\theta_{n+1}|\theta_h)$. However, because a $N(0,10^6)$ prior incorporates no prior information about θ_{n+1} in the analysis of each simulated data set, $\Pr(\mathcal{H}_A|D_h) = 0.5$ and hence $\text{pow}(m_{n+1}|D_h) = 0$ when $m_{n+1} = 0$.

The low dimensionality of the Bayesian nonhierarchical covariate-free model allowed the trapezoidal method to be used for integration.[16] The fast rate of computation allowed precise evaluations to be performed for all values of $m_{n+1} \in \{0, 1, \ldots, 500\}$, enabling the exact form of the power function to be determined.

For other models, stochastic optimization was performed using data from two independent simulations. The simulation was conducted with 12 distinct values of m_{n+1} ranging from 5 to 800 (see Table 2). The design points for the simulation experiment were skewed to the right because the power function is more sensitive over small values of m_{n+1}. Precise estimates of the power were obtained by performing 1000 simulations at each value of m_{n+1}, implying a total of $S = 12000$ simulations. The standard error of estimates of power and sample size were evaluated using a nonparametric bootstrap involving 250 bootstrap samples in which outcomes were sampled with replacement within the groups of 1000 simulated outcomes associated with each distinct value of m_{n+1}.

Table 2. Power of Noninferiority Test at Fixed Sample Sizes by Parameter Weighting Distribution, for Different Methods of Computation

Sample Size (m_f)	Current Method				Bayesian: Nonhierarchical Covariate-Free				Bayesian: Hierarchical Covariate-Free				Bayesian: Hierarchical Covariate-Adjusted			
	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4
5	0.065	0.059	—	—	0.576	0.476	0.754	0.600	0.570	0.482	0.750	0.593	0.554	0.469	0.740	0.549
10	0.099	0.074	—	—	0.332	0.227	0.575	0.375	0.318	0.204	0.565	0.373	0.352	0.249	0.570	0.399
20	0.168	0.099	—	—	0.368	0.216	0.692	0.428	0.383	0.221	0.698	0.442	0.372	0.250	0.664	0.412
35	0.273	0.133	—	—	0.497	0.271	0.841	0.552	0.455	0.251	0.789	0.514	0.448	0.255	0.772	0.479
50	0.375	0.166	—	—	0.576	0.297	0.900	0.612	0.492	0.236	0.869	0.570	0.503	0.275	0.846	0.554
55	0.407	0.176	—	—	0.482	0.213	0.867	0.549	0.527	0.235	0.864	0.590	0.537	0.268	0.872	0.564
80	0.557	0.228	—	—	0.677	0.322	0.948	0.669	0.653	0.285	0.940	0.635	0.654	0.364	0.928	0.660
100	0.656	0.268	—	—	0.756	0.367	0.969	0.708	0.730	0.345	0.965	0.701	0.709	0.373	0.961	0.688
200	0.922	0.452	—	—	0.930	0.503	0.997	0.784	0.900	0.468	0.989	0.795	0.897	0.488	0.993	0.777
350	0.994	0.664	—	—	0.992	0.690	1.000	0.837	0.989	0.640	1.000	0.849	0.990	0.636	0.999	0.830
550	1.000	0.838	—	—	0.999	0.814	1.000	0.863	1.000	0.773	1.000	0.867	1.000	0.773	1.000	0.871
800	1.000	0.940	—	—	1.000	0.921	1.000	0.889	1.000	0.877	1.000	0.882	1.000	0.895	1.000	0.881

The parameter weighting distributions are defined on $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h = 0\}(w_1)$, $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h = 0.315\}(w_2)$, $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h \leq 0\}(w_3)$, and $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h \leq 0.63\}(w_4)$. Results for the current method are only reported for w_1 and w_2, indicative of current practice. The historical control comprises 5806 patients with $\hat{p}_h = 0.104$.

RESULTS

True Power Function

Figure 1 is a plot of the power function for the Bayesian nonhierarchical covariate-free model with weights generated by w_4. The power function does not exhibit the usual monotone increasing trend, but rather consists of clusters in which power decreases with sample size. The average power associated with each cluster generally increases with sample size.

The form of the power function in Fig. 1 is due to the discrete nature of the outcome (Appendix C, Result 3). The logistic regression model in Eq. (12) can be thought of as approximating a smooth version of the true power function, such as that represented by the mean values of power and m_{n+1} for each cluster of points. Because of the irregular behavior of the power function for small values of m_{n+1}, simulated data with $m^s_{n+1} \leq 20$ for w_i $(i \neq 2)$, and $m^s_{n+1} \leq 55$ for w_2, were not used to fit Eq. (12).

Comparison of Methods for Computing Power

Table 2 displays empirical estimates of power using the current method, and assuming each of the three Bayesian models (nonhierarchical covariate-free, hierarchical covariate-free, and hierarchical covariate-adjusted) for the analysis. The standard errors (not shown) of estimates based on stochastic optimization were generally lower than those for independent calculations, as information is shared between design points. No results are reported for w_3 and w_4 under the current method, because this method is to substitute estimates for parameters rather than averaging over them.

When $m_{n+1} \leq 100$ the Bayesian methods yield higher values of power than the currently used calculation [Eq. (13)]. However, as m_{n+1} becomes large Bayesian evaluations are more conservative than the current method. For example, under w_2, the current method estimates the power at $m_{n+1} = 50$ and $m_{n+1} = 800$ to be 0.166 and 0.940, while the Bayesian nonhierarchical covariate-free method yields estimates 0.297 and 0.921, respectively.

Among the Bayesian models, the hierarchical covariate-free model tended to yield lower values of power than the nonhierarchical covariate-free model. For example, at $m_{n+1} = 50$ and based on w_2, power is estimated as 0.869 and 0.900 under the two models, respectively. Accounting for the correlation among observations within the historical trials inflates the variability of the overall TLR rate and thus reduces power. Because of the underlying form of the power function (see the section "True Power Function"), covariate adjustment led to higher values of power for some but not all values of m_{n+1}.

Comparison of the results for different parameter weighting distributions were as expected, with higher values obtained under w_1 than w_2, and higher values obtained under w_3 than either w_1 or w_4. The power function tended to

have a steeper gradient under w_1 and w_2 than under w_3 and w_4, respectively, indicating that power is more sensitive to m_{n+1} when the effect size $\theta_{n+1} - \bar{\theta}_h$ is fixed.

Estimates of Optimal Sample Size

Point estimates of the optimal sample sizes based on Eq. (12) are obtained by tracing the nominal level of power onto the m_{n+1}-axis as depicted in Fig. 2. The closeness of the fitted curve to the empirical estimates of power over the region $m_{n+1} > 50$, suggests that Eq. (12) is a good model of the true power function. The power function under w_4 appears to have an asymptote around 0.9, a characteristic predicted by the theoretical results in Appendix B. Because Eq. (12) tends to 1 as $m_{n+1} \to \infty$, over-prediction of power is likely to occur at large m_{n+1}.

From Fig. 2, the optimal sample size based on the Bayesian hierarchical covariate-adjusted model when $1 - \gamma = 0.80$ is 136 subjects under w_1, and 585 subjects for w_2. The standard errors associated with these evaluations, 3.054 and

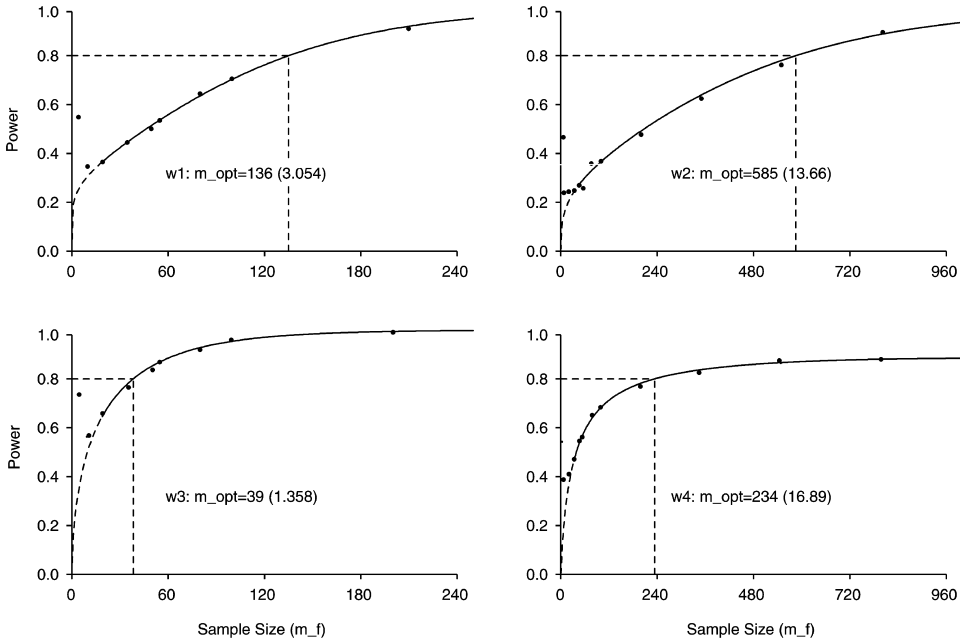

Figure 2. Power curve for the Bayesian hierarchical covariate-adjusted model for four uniform parameter weighting distributions: $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h = 0\}$ (w_1), $\Theta_{n+1} = \{\theta : \theta_{n+1} - \bar{\theta}_h = 0.315\}$ (w_2), $\Theta_{n+1} = \{\theta : -1.5 \le \theta_{n+1} - \bar{\theta}_h \le 0\}$ (w_3), $\Theta_{n+1} = \{\theta : -0.87 \le \theta_{n+1} - \bar{\theta}_h \le 0.63\}$ (w_4). The historical control comprised of 5806 patients with $\hat{p}_h = 0.104$. The optimal sample size at power 0.8 is displayed in each case. The power function has a broken line for those values of m_{n+1} where the true power function is highly disjointed.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

13.66, respectively, are due to lack-of-fit of the model given in Eq. (12), and the use of a finite number of simulations.

These Bayesian evaluations compare with 139 and 495 for the current method [Eq. (13)]. Clearly, if the current method for sample size determination is used, the new trial will be under-powered under w_2.

DISCUSSION

Tests of new medical devices are frequently performed using retrospective control groups formed from trials of approved devices. Tests such as those described here are also used for mechanical and biological prosthetic heart valves, and other medical devices. The stent data include heterogeneous covariate distributions among stent trial participants. We proposed a methodology that accounted for covariate effects, and the uncertainty in parameter values. This extends previous methods developed by Spiegelhalter and Freedman,[12] Weiss,[13] Gould,[17] and Thall and Simon[18] in various ways.

Our general methodology can be adapted in several important directions. Parameter weighting distributions incorporating polynomial, logarithmic, exponential, or other functions may be substituted for Eq. (11). Moreover, criteria other than power may be used for sample size determination. For example, the probability of choosing correctly between $\mathcal{H}_0$ or $\mathcal{H}_A$, or a utility function that offsets the cost of making an incorrect decision with the cost of sampling (see, for example, Refs. [5,19]), could serve as the basis of the calculations.

The strongest reasons for adopting our approach are practical rather than theoretical. Current methods for determining sample sizes often assume analyses that bear little resemblance to analyses that are appropriate for historically controlled trials. This makes the resulting calculations unreliable and can lead to trials that are either seriously under- or over-powered. Although analytic results are difficult to derive when intended analyses are complex, convenient solutions can be obtained through the use of large-scale numerical integration.

APPENDIX A: POSTERIOR INFERENCES

Two different Markov Chain Monte Carlo (MCMC) computations are required to evaluate power for the Bayesian models. The first is used to draw (θ_h, ξ) from $f(\theta_h, \xi|D_h)$, from which model parameters and data for a hypothetical new trial are generated. The second evaluates $\Pr(\mathcal{H}_A|y^s_{n+1}, X^s_{n+1}, D_h)$ and involves modeling both the new and historical data $(y^s_{n+1}, X^s_{n+1}, D_h)$.

The Markov chains we employ are based on the Gibbs Sampler,[20,21] and use Metropolis–Hastings[22,23] steps to draw from conditional posterior densities that are only known up to proportionality. The conditional posterior distributions for the Bayesian-hierarchical covariate-adjusted model fitted to

$(y_{n+1}, X_{n+1}, D_{\mathrm{h}})$ are provided. The conditional posterior distributions required to fit the same model to just the historical data D_{h} are special cases.

$$(\tau^2|\cdot) \sim \mathrm{IG}\left[u_0 + \frac{n}{2}, \left\{\frac{1}{v_0} + \frac{1}{2}\sum_{i=1}^{n}(\theta_i - \mu)^2\right\}^{-1}\right],$$

$$(\mu|\cdot) \sim \mathrm{N}\left(\frac{s_0^2\sum_{i=1}^{n}\theta_i + \tau^2 a_0}{ns_0^2 + \tau^2}, \frac{s_0^2\tau^2}{ns_0^2 + \tau^2}\right),$$

$$(\beta|\cdot) \propto \prod_{i\in\{1,\ldots,n+1\}} \prod_{j=1}^{m_i} \exp\{(\theta_i + \beta^T x_{ij})y_{ij}\}$$

$$\times \{1 + \exp(\theta_i + \beta^T x_{ij})\}^{-1} \exp\left\{-\frac{1}{2}(\beta - b_0)^T S_0^{-1}(\beta - b_0)\right\},$$

$$(\theta_i|\cdot) \propto \prod_{j=1}^{m_i} \exp\{(\theta_i + \beta^T x_{ij})y_{ij}\}$$

$$\times \{1 + \exp(\theta_i + \beta^T x_{ij})\}^{-1} \exp\left\{-\frac{1}{2\tau^2}(\theta_i - \mu)^2\right\},$$

$$(\theta_{n+1}|\cdot) \propto \prod_{j=1}^{m_{n+1}} \exp\{(\theta_{n+1} + \beta^T x_{(n+1)j})y_{(n+1)j}\}$$

$$\times \{1 + \exp(\theta_{n+1} + \beta^T x_{(n+1)j})\}^{-1} \exp\left\{-\frac{1}{2t_0^2}(\theta_{n+1} - m_0)^2\right\},$$

where p is the size of β.

Metropolis–Hastings steps are required to draw samples of β, θ_{h}, and θ_{n+1}. Candidate samples of β are drawn from multivariate T proposal densities, while samples of θ_{i} $(i = 1, \ldots, n+1)$ are drawn from Student's T distribution. Degrees of freedom parameters were set to 3 to ensure that the tails of the proposal density were thicker than those of the true posterior. In turn, this helps to ensure the chain is geometrically ergodic,[24] and hence rapid convergence to the stationary distribution. The scale of the covariance matrix was chosen so that the acceptance rate of proposed variates was in the range $[0.3, 0.5]$, allowing efficient exploration of the parameter space.

An essential component of MCMC computation is ensuring that the chain converges to the posterior distribution of the model parameters. Convergence of the Markov chain was examined using the CODA (Convergence Diagnosis and Output Analysis Software for Gibbs sampling output) software package.[25] The behavior of the chain was monitored using trace plots[26] of the sequence of draws for each

parameter for different starting values and random number seeds. Convergence was formally checked using convergence diagnostics available in CODA.

Final computations were performed with a burn-in of 2000 iterations followed by 10,000 additional iterations for each simulated data set. These values far exceeded the values of 1000 and 2000, respectively, suggested as appropriate by the Gelman and Rubin scale reduction factors[27] and the Raftery and Lewis dependence factors[28] in CODA, and our own empirical studies. Trace plots indicated that although serial correlation was slightly higher for parameters sampled using Metropolis–Hastings steps than for other parameters, the Markov chain rapidly explored the parameter space. Chains quickly converged from different starting values and random number seeds. Approximately 48 hr were needed to analyze 1000 simulated trials under the Bayesian hierarchical covariate-adjusted model on a Sun Ultra-Sparc computer (Sun Microsystems, Palo Alto, CA, USA). Code for performing the computations was written in the C programming language.

APPENDIX B: ASYMPTOTIC RESULTS

The Bayesian nonhierarchical covariate-free model is assumed in this Appendix. We derive a normal approximation of $\theta_{n+1} - \bar{\theta}_h$ from which power is determined analytically, leading to two asymptotic results.

The posterior distribution of $\bar{\theta}_h$ is approximately $N(\lambda_h, \sigma_h^2)$, where $\lambda_h = E[\bar{\theta}_h|D_h]$ is the mean and $\sigma_h^2 = \text{var}[\bar{\theta}_h|D_h]$ is the variance. The empirical logit $\hat{\theta}_{n+1} = \log\{y_{(n+1)}/(m_{n+1} - y_{(n+1)})\}$, where $y_{(n+1)} = \sum_{j=1}^{m_{n+1}} y_{(n+1)j}$, is the maximum likelihood estimator of θ_{n+1}. It follows from the asymptotic normality of maximum likelihood estimates for generalized linear models[29] that $\hat{\theta}_{n+1}$ is asymptotically normally distributed with mean θ_{n+1} and variance

$$\frac{\sigma_{n+1}^2}{m_{n+1}} = \frac{\{1 + \exp(\theta_{n+1})\}^2}{m_{n+1}\exp(\theta_{n+1})}.$$

The prior assumed for θ_{n+1} is approximately locally uniform; therefore, we may proceed as if $f(\theta_{n+1}) \propto 1$. Temporarily treating σ_{n+1}^2 as a known constant (ignoring the dependence on θ_{n+1}) it follows that $f(\theta_{n+1}|y_{n+1}) = N(\hat{\theta}_{n+1}, \sigma_{n+1}^2)$, and therefore $\theta_{n+1} - \bar{\theta}_h|\hat{\theta}_{n+1}, x_{(n+1)j} \simeq N(\hat{\theta}_{n+1} - \lambda_h, \sigma_{n+1}^2/m_{n+1} + \sigma_h^2)$. The critical region for the level-α test of $\mathcal{H}_0 : \theta_{n+1} \geq \bar{\theta}_h + \delta_0$ vs. $\mathcal{H}_A : \theta_{n+1} < \bar{\theta}_h + \delta_0$ is $\hat{\theta}_{n+1} < \delta_0 + \lambda_h - z_\alpha(\sigma_{n+1}^2/m_{n+1} + \sigma_h^2)^{1/2}$. The probability of this region assuming $\hat{\theta}_{n+1} \sim N(\bar{\theta}_h + \delta_A, \sigma_{n+1}^2/m_{n+1})$ is given by

$$\widehat{\text{pow}}(m_{n+1}, \delta_A, \bar{\theta}_h|D_h) = \Phi\left\{\frac{\delta_0 - \delta_A + \lambda_h - \bar{\theta}_h}{(\sigma_{n+1}^2/m_{n+1})^{1/2}} - z_\alpha\left(1 + \frac{m_{n+1}\sigma_h^2}{\sigma_{n+1}^2}\right)^{1/2}\right\}. \quad \text{(B1)}$$

Result 1. *The probability in Eq. (B1) converges to 1, 1/2, and 0 as* $m_{n+1} \to \infty$ *if*

$$\bar{\theta}_h \arg \lambda_h + \delta_0 - \delta_A - z_\alpha \sigma_h, \tag{B2}$$

where arg is " $<$ *", "* $=$ *," and "* $>$ *," respectively.*

Proof. As Φ is a monotone function of m_{n+1} we need only consider the behavior of the argument. This can be written in the form $g(m_{n+1}) = \{k_2(1/m_{n+1} + k_3)^{1/2} - k_1\}m_{n+1}^{1/2}$, where $k_1 = (\bar{\theta}_h + \delta_A - \delta_0 - \lambda_h)/\sigma_{n+1}$, $k_2 = -z_\alpha$ and $k_3 = \sigma_h^2/\sigma_{n+1}^2$. The expression in Eq. (B2) is equivalent to $k_1 \arg k_2 k_3^{1/2}$. When $k_1 < k_2 k_3^{1/2}$ it is clear that $k_2(1/m_{n+1} + k_3)^{1/2} - k_1 \to l > 0$, and therefore that $g(m_{n+1}) \to \infty$. Hence, when arg is "$<$," Eq. (B1) converges to 1 as $m_{n+1} \to \infty$. When $k_1 > k_2 k_3^{1/2}$, an analogous argument reveals that Eq. (B1) converges to 0 as $m_{n+1} \to \infty$. The definition of a limit of a continuous function can be used to show that $g(m_{n+1}) \to 0$, and hence that Eq. (B1) converges to 1/2, when $k_1 = k_2 k_3^{1/2}$.□

Result 2. *The expected value of Eq. (B1) with respect to* $\bar{\theta}_h | \lambda_h, \sigma_h^2 \sim N(\lambda_h, \sigma_h^2)$, *an approximation of the prior for* $\bar{\theta}_h$ *given* D_h, *converges to*

$$\Phi\left\{\frac{\delta_0 - \delta_A}{\sigma_h} - z_\alpha\right\} \tag{B3}$$

as $m_{n+1} \to \infty$.

Proof. From Result 1, the power probability converges to 1, 1/2, or 0 depending on the value of $\hat{\theta}_h$. As the limiting values are finite it follows that the limit of the expectation can be expressed as the expectation of the limit. Since $\Pr(\bar{\theta}_h = \lambda_h + \delta_0 - \delta_A - z_\alpha \sigma_h) = 0$, it follows that the expected value of Eq. (B1) tends to $\Pr(\bar{\theta}_h < \lambda_h + \delta_0 - \delta_A - z_\alpha \sigma_h)$. By our assumption $\bar{\theta}_h \sim N(\lambda_h, \sigma_h^2)$ and the expression in Eq. (B3) is obtained. □

APPENDIX C: POWER FUNCTION FOR BINARY OUTCOMES

Result 3. *The power function for a binary outcome consists of clusters in which power decreases with sample size.*

Justification. The sufficient statistics for the Bayesian nonhierarchical covariate-free model are $(\sum_{i=1}^{m_{n+1}} y_{(n+1)i}, m_{n+1})$. Therefore, $\Pr(\mathcal{H}_A | y_{n+1}, X_{n+1}, D_h)$ takes $m_{n+1} + 1$ distinct values corresponding to the number of failures out of m_{n+1}. Let $y_{(n+1).}^c = \min\{\sum_{i=1}^{m_{n+1}} y_{(n+1)i} : \Pr(\mathcal{H}_A | y_{n+1}, X_{n+1}, D_h) \geq 1 - \alpha\}$ denote the critical point at which $\mathcal{H}_A$ is concluded. Due to the discrete nature of the response, $y_{(n+1)}^c$ can be the same for successive values of m_{n+1}, in which case

power is higher at the smaller value because the critical point is easier to attain. For example, if the critical point associated with $m_{n+1} = 5$ and $m_{n+1} = 6$ is $y^c_{(n+1)} = 0$ (no occurrences of TLR), then because 0 failures out of 5 is more likely than 0 failures out of 6, the power of the test is higher at $m_{n+1} = 5$. Moreover power increases only when $y^c_{(n+1)}$, increases with m_{n+1}.

ACKNOWLEDGMENTS

The authors thank the Advanced Medical Manufacturers Association (Washington, DC), a nonprofit consortium of health care product manufacturers for support; and Gene Penello (Division of Biostatistics, Food and Drug Administration, Rockville, MD) for valuable discussions. In addition, we thank two anonymous referees, the editor, and an associate editor for providing comments that greatly improved the manuscript.

REFERENCES

1. Self, S.G.; Mauritsen, R.H. Power/Sample Size Calculations for Generalized Linear Models. Biometrics **1988**, *44*, 79–86.
2. Whittemore, A.S. Sample Size for Logistic Regression with Small Response Probability. J. Am. Stat. Assoc. **1981**, *76*, 27–32.
3. Hsieh, F.Y.; Bloch, D.A.; Larsen, M.D. A Simple Method of Sample Size Calculation for Linear and Logistic Regression. Stat. Med. **1998**, *17*, 1623–1634.
4. Signorini, D.F. Sample Size for Poisson Regression. Biometrika **1991**, *78*, 446–450.
5. Müller, P.; Parmigiani, G. Optimal Design via Curve Fitting of Monte-Carlo Experiments. J. Am. Stat. Assoc. **1995**, *90*, 1322–1330.
6. Kereiakes, D.J.; Linnemeier, T.J.; Baim, D.S.; Kuntz, R.E.; O'Shaghnessy, C.; Hermiller, J.; Fink, S.; Lansky, A.; Nishimura, N.; Broderick, T.M.; Popma, J.J. Procedural and Late Outcomes Following Multi-Link Duet Coronary Stent Deployment. Am. J. Cardiol. **1999**, *84*, 1385–1390.
7. O'Malley, A.J.; Normand, S.-L.T.; Kuntz, R.E. Estimation of the Objective Performance Criterion for Medical Device Trials: Coronary Artery Stenting. Stat. Med. **2002**, in press.
8. Blackwelder, W.C. Proving the Null Hypothesis in Clinical Trials. Control. Clin. Trials **1982**, *3*, 345–353.
9. Gersh, B.J.; Fisher, L.D.; Schaff, H.V.; Rahimtoola, S.H.; Reeder, G.S.; Frater, R.W.M.; McGoon, D.C. Issues Concerning the Clinical Evaluation of New Prosthetic Valves. J. Thorac. Cardiovasc. Surg. **1986**, *91*, 460–466.
10. Grunkemeier, G.L. Will Randomized Trials Detect Random Valve Failure? Reflections on a Recent FDA Workshop. J. Heart Valve Dis. **1993**, *2*, 424–429.
11. Pocock, S.J. The Combination of Randomized and Historical Controls in Clinical Trials. J. Chronic Dis. **1976**, *29*, 175–188.
12. Spiegelhalter, D.J.; Freedman, L.S. A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion. Stat. Med. **1986**, *5*, 1–13.
13. Weiss, R. Bayesian Sample Size Calculations for Hypothesis Testing. Statistician **1997**, *46*, 185–191.

14. Atkinson, D. Assessing Nonsampling Errors in Survey Data Through Random Intercept Models. American Statistical Association Proceedings of the Section on Survey Research Methods, **1997**, 401–406.
15. Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. **1979**, *7*, 1–26.
16. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th Ed.; Dover Publications: New York, 1972.
17. Gould, A.L. Sample Sizes for Event Rate Equivalence Trials Using Prior Information. Stat. Med. **1993**, *12*, 2009–2023.
18. Thall, P.F.; Simon, R.M. Incorporating Historical Control Data in Planning Phase II Clinical Trials. Stat. Med. **1990**, *9*, 215–228.
19. Lindley, D.V. The Choice of Sample Size. Statistician **1997**, *46*, 129–138.
20. Geman, S.; Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans. Pattern Anal. Machine Intell. **1984**, *6*, 721–741.
21. Gelfand, A.E.; Smith, A.F.M. Sampling Based Approaches to Calculating Marginal Densities. J. Am. Stat. Assoc. **1990**, *85*, 398–409.
22. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equations of State Calculations by Fast Computing Machines. J. Chem. Phys. **1953**, *21*, 1087–1091.
23. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika **1970**, *57*, 97–109.
24. Jones, G.L.; Hobert, J.P. Honest Exploration of Intractable Probability Distributions Via Markov Chain Monte Carlo. Stat. Sci. **2001**, *16*, 312–334.
25. Best, N.; Cowles, M.K.; Vines, K. *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*; MRC Biostatistics Unit, Institute of Public Health: Cambridge, UK, 1995.
26. Hellmich, M.; Abrams, K.R.; Jones, D.R.; Lambert, P.C. A Bayesian Approach to a General Regression Model for ROC Curves. Acad. Radiol. **1998**, *18*, 436–443.
27. Gelman, A.; Rubin, D.B. Inference from Iterative Simulation Using Multiple Sequences. Stat. Sci. **1992**, *7*, 457–472.
28. Raftery, A.L.; Lewis, S.M. How Many Iterations in the Gibbs Sample? In *Bayesian Statistics*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University Press: Oxford, 1992; 763–774.
29. Agresti, A. *Categorical Data Analysis*; John Wiley and Sons: New York, 1990; 427–428.

Received January 2002
Revised July 2002
Accepted July 2002