Bayesian Sample Size Determination for Case-Control Studies

Cyr Emile M'LAN, Lawrence JOSEPH, and David B. WOLFSON

Case-control studies are among the most commonly used means of assessing association between exposure and outcome. Sample size determination and the optimal control-to-case ratio are vital to the design of such studies. In this article we investigate Bayesian sample size determination and the control-to-case ratio for case-control studies, when interval estimation is the goal of the eventual statistical analysis. In certain cases we are able to derive approximate closed-form sample size formulas. We also describe two Monte Carlo methods, each of which provides a unified approach to the sample size problem, because they may be applied to a wide range of interval-based criteria. We compare the accuracy of the different methods. We also extend our methods to include cross-sectional designs and designs for gene–environment interaction studies.

KEY WORDS: Case-control studies; Cross-sectional studies; Gene–environment interaction studies; Highest posterior density intervals; Odds-ratio; Optimal control-to-case ratio; Sample size determination.

1. INTRODUCTION

The case-control study is an observational approach to studying exposure–disease relationships (Schlesselman 1982). Traditional case-control studies compare the exposures in a sample of n_1 diseased cases that occur during an accrual time period, with the exposure in a sample of n_0 control subjects alive and free of disease and from the same population as the cases. The investigator must retrospectively collect information on exposure status for each case and each control. Let D and E represent the disease and exposure status. Let $p_1 = \Pr(E = 1|D = 1)$ and $p_0 = \Pr(E = 1|D = 0)$ denote the conditional exposure probabilities among case and control subjects. A common measure of disease–exposure association is the exposure odds ratio, $\psi_e = \frac{p_1(1-p_0)}{p_0(1-p_1)}$. Because the exposure odds ratio is equal to the disease odds ratio (Cornfield 1951), defined by $\frac{p_1'(1-p_0')}{p_0'(1-p_1')}$, where $p_1' = \Pr(D = 1|E = 1)$ and $p_0' = \Pr(D = 1|E = 0)$, values of $\psi_e > 1$ (resp. $\psi_e < 1$) indicate that exposure $E = 1$ is associated with increased (resp. decreased) risk of disease. When the disease is rare, the disease odds ratio closely approximates the relative risk, defined by $\frac{p_1'}{p_0'}$, a more direct measure of exposure effect.

Before carrying out a case-control study, among the most important design issues is determining the required sample size and the proper control-to-case ratio. The extensive literature on the sample size problem for case-control studies has been reviewed by Wickramaratne (1995), with the greatest attention focused on power-based sample size criteria (Schlesselman 1982). A trend toward the use of confidence intervals rather than p values in the statistical analysis of medical studies (Gardner and Altman 1986) has shifted attention to interval-based sample size criteria, however. In general, it is desirable that sample size criteria be consistent with the methods used in the eventual analysis. O'Neill (1984) implemented an interval-based sample size criterion for case-control studies, and Satten and Kupper (1990) proposed a similar criterion based on tolerance probabilities. Some (e.g., Nam and Fears 1992) have extended the sample size problem for case-control studies to include designs beyond simple 2×2 tables, including stratified studies and cost considerations. All of these authors have restricted their attention to frequentist methods.

The most commonly used frequentist sample size methods for case-control studies have several drawbacks. They rely on the adequacy of the normal approximation to the binomial distribution, requiring special refinements when the rare disease assumption is valid (Lemeshow, Hosmer, and Stewart 1981). All frequentist sample size formulas depend on accurate point estimates of p_0 and p_1 and can be sensitive to minor changes in these estimates. Also, by using only point estimates, these procedures do not fully use all available prior information, which includes the degree of uncertainty about the values of p_0 and p_1.

In the Bayesian framework, sample size determination is enhanced through the specification of a prior distribution for the unknown parameter vector (p_0, p_1). Furthermore, by averaging the potential data with respect to its marginal distribution, the methodology also takes into consideration the stochastic nature of the as-yet unobserved data. The approach also does not depend on a normal approximation to the posterior distribution. Of course, if the prior-likelihood-posterior paradigm were invoked at the design stage, then it is anticipated that the same paradigm would be used at the analysis stage, in keeping with the earlier-stated principle of design/analysis consistency.

Various Bayesian sample size criteria have been proposed and applied to a range of parameter estimation problems, including multinomial probabilities (Adcock 1987, 1988, 1993), single- and two-sample normal means (Joseph and Bélisle 1997), single binomial proportions (Pham-Gia and Turkkan 1992; Joseph, Wolfson, and du Berger 1995), and the difference between two binomial proportions (Joseph, du Berger, and Bélisle 1997). These criteria have been summarized by Adcock (1997). Wang and Gelfand (2002) reviewed Bayesian Monte Carlo methods for sample size determination.

Whereas Bayesian methods have been widely applied to the analysis of case-control data including matched studies (Ghosh and Chen 2002) and to studies with misclassified data (Müller and Roeder 1997), very little has been done in Bayesian optimal design for case-control studies. In Bayesian design problems, one must consider all possible datasets that can arise,

Cyr Emile M'Lan is Assistant Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 (E-mail: *mlan@merlot.stat.uconn.edu*). Lawrence Joseph is Associate Professor, Department of Epidemiology and Biostatistics (E-mail: *lawrence.joseph@mcgill.ca*), and David B. Wolfson is Professor, Department of Mathematics and Statistics (E-mail: *david@math.mcgill.ca*), McGill University, Montreal, Quebec, Canada H3G1A4.

© 2006 American Statistical Association
Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000001023

whereas in analysis problems one typically considers just a single dataset. Thus it is not surprising that sample size methods tend to lag behind the corresponding analysis methods. Nevertheless, De Santis and Perone Pacifico (2001) and De Santis, Perone, and Sambucini (2004) were the first to systematically consider Bayesian sample size problems for case-control studies. They considered both interval-based and test-based criteria and examined the optimal control-to-case ratio. Our work differs from and extends the work of De Santis and colleagues in several respects. We introduce new sample size criteria and show how two Monte Carlo methods provide a unified approach to the sample size problem when applied to a wide range of interval-based criteria. Moreover, in certain cases we are able to derive approximate closed-form sample size formulas, and all of our methods rely on more efficient highest posterior density (HPD) intervals. We further consider a wider range of problems beyond those arising from 2×2 tables, including higher-order tables that arise from design problems for gene–environment interaction studies. Our methods are also shown to apply to cross-sectional sampling designs.

The article is organized as follows. In Section 2 we review previous Bayesian sample size criteria and propose several new criteria. To limit the length of the article and for technical reasons, we focus on the average length criterion (ALC). Section 3 begins by developing an approximate closed-form expression for the optimal sample size for the ALC. Pointing out the difficulties in deriving analogous closed form expressions for the other criteria, we introduce a straightforward Monte Carlo approach, the accuracy and efficiency of which we then improve by introducing a regression component. Section 3 concludes with an illustrative example. Section 4 discusses extensions to cross-sectional (multinomial) sampling and studies of gene–environment interactions. Section 5 covers Bayesian decision-theoretic sample size criteria, and Section 6 provides some concluding remarks.

2. BAYESIAN SAMPLE SIZE CRITERIA

Let $\theta \in \Theta$ be a scalar parameter distributed a priori as $f(\theta)$, and let $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathcal{X}_n$ be n iid realizations of a random variable X distributed as $p(x|\theta), x \in \mathcal{X}$, given θ. Let $p_{X_n}(\mathbf{x}_n|n) = \int_\Theta p(\mathbf{x}_n|\theta, n) f(\theta)\, d\theta$ be the marginal distribution. Denote the posterior density of θ by $f_\theta(\theta|\mathbf{x}_n, n)$. Let $\mathrm{HPD}_L(\mathbf{x}_n, n, l) = (u, v)$, $u < v$, be an HPD interval for θ of length l, and let $\mathrm{HPD}_C(\mathbf{x}_n, n, 1-\alpha)$ be an HPD interval for θ of posterior coverage $1-\alpha$. Define $l^\star_{1-\alpha}(\mathbf{x}_n|n)$ and $\alpha^\star_l(\mathbf{x}_n|n) = \int_{\mathrm{HPD}_L(\mathbf{x}_n,n,l)} f(\theta|\mathbf{x}_n, n)\, d\theta$ to be the length and the posterior coverage of an HPD interval of coverage $1-\alpha$ and of length l.

Letting k be a positive integer, we define the kth average coverage criterion (ACC_k) sample size to be the minimum n such that

$$\left(\int_{\mathcal{X}_n} \{\alpha^\star_l(\mathbf{x}_n|n)\}^k p_{X_n}(\mathbf{x}_n|n)\, d\mathbf{x}_n \right)^{1/k} \geq 1-\alpha. \tag{1}$$

This yields the smallest n for which HPD intervals of length l provide an average posterior coverage of at least $1-\alpha$, where the average is with respect to the marginal distribution under the L_k norm, that is, the kth marginal moment of α^*_l.

Conversely, the kth average length criterion, ALC_k, fixes the desired posterior coverage at $1-\alpha$. The lengths of all HPD intervals are averaged with respect to the marginal distribution $p_{X_n}(\mathbf{x}_n|n)$ under the L_k norm. One then seeks the minimum n such that

$$\left(\int_{\mathcal{X}_n} \{l^\star_{1-\alpha}(\mathbf{x}_n|n)\}^k p_{X_n}(\mathbf{x}_n|n)\, d\mathbf{x}_n \right)^{1/k} \leq l. \tag{2}$$

The ACC_k and ALC_k extend and include, as special cases, the $\mathrm{ACC} = \mathrm{ACC}_1$ and $\mathrm{ALC} = \mathrm{ALC}_1$, first proposed by Joseph et al. (1995). The choices $k = 1, 2$, and $+\infty$ are of particular interest, because they lead to criteria that are easy to interpret. In the sequel we discuss these important special cases.

The ACC and ALC guarantee their posterior coverages and lengths only on average with respect to the marginal distribution. In contrast, the "worst outcome" criterion (WOC; Joseph et al. 1995) and its modified version, MWOC (Joseph et al. 1997) are preferred when one desires a more conservative sample size, which guarantees the length and posterior coverage over all anticipated data $\mathbf{x}_n$ or over a subset $\mathcal{S}_n$ of $\mathcal{X}_n$. For example, $\mathcal{S}_n$ could be a $100(1-\gamma)\%$ credible region according to the marginal distribution, p_{X_n}. Fixing the coverage, this conservative criterion would then seek the minimum n such that

$$\inf_{\mathbf{x}_n \in \mathcal{S}_n} \alpha^\star_l(\mathbf{x}_n|n) \geq 1-\alpha \tag{3}$$

or, equivalently,

$$\sup_{\mathbf{x}_n \in \mathcal{S}_n} l^\star_{1-\alpha}(\mathbf{x}_n|n) \leq l.$$

We introduce two new criteria, the median coverage criterion, (MCC) and the median length criterion (MLC), which lead to the minimum n such that

$$\operatorname*{med}_{\mathbf{x}_n \in \mathcal{X}_n} \alpha^\star_l(\mathbf{x}_n|n) \geq 1-\alpha \tag{4}$$

and

$$\operatorname*{med}_{\mathbf{x}_n \in \mathcal{X}_n} l^\star_{1-\alpha}(\mathbf{x}_n|n) \leq l. \tag{5}$$

Rather than depending on HPD intervals, all of the foregoing criteria can also be defined in terms of easier-to-compute equal-tailed intervals. Pham-Gia and Turkkan (1992) proposed two criteria, PGT-(i) and PGT-(ii), based on posterior variances, which are equivalent, to the WOC and ALC_2 defined earlier as the sample size increases to infinity. When $\mathcal{X}_n$ is a discrete set, it can also be shown that $\mathrm{ALC}_\infty = \mathrm{WOC}$.

3. BAYESIAN SAMPLE SIZE METHODS FOR CASE–CONTROL STUDIES

As discussed earlier, we focus mainly on the ALC_k, pointing out where there is general applicability of our methods to the ACC_k, WOC, MWOC, MLC, and MCC. The problem is as follows. Determine the minimum total sample size, $n = n_0 + n_1$, where n_0 and n_1 are the number of controls and cases to be sampled. Let the control-to-case ratio be $g = n_0/n_1$. Concurrent to the sample size problem, find g.

3.1 Bayesian Modeling of Case-Control Studies

We use the convenient notation $T = (a, n_1 - a, c, n_0 - c)$ to notate data arising from a typical case-control study, as il-

Table 1. Generic 2 × 2 Table for Exposure–Disease Outcomes

	$E = 1$	$E = 0$	*Total*
$D = 1$	a	$n_1 - a$	n_1
$D = 0$	c	$n_0 - c$	n_0

lustrated in Table 1. The numbers exposed among the cases and the controls, are assumed to be independently binomially distributed conditional on (p_0, p_1), with $a \sim \text{Bin}(n_1, p_1)$ for the cases and $c \sim \text{Bin}(n_0, p_0)$ for the controls. Most Bayesian analyses of case-control studies (e.g., Zelen and Parker 1986; Nurminen and Mutanen 1987; Marshall 1988; Carlin 1992; Ashby, Hutton, and McGee 1993; Hashemi, Nandram, and Goldberg 1997) assume that p_1 and p_0 are a priori independent, with $p_1 \sim \text{Be}(a', b')$ and $p_0 \sim \text{Be}(c', d')$, where $\text{Be}(a, b)$ represents the beta distribution with parameters (a, b). We also follow this practice.

Let $T' = (a', b', c', d')$ denote the four prior parameters. The combination of the prior and likelihood tables, T' and T, leads to a posterior table, $T'' = (a'', b'', c'', d'')$, where $a'' = a + a'$, $b'' = n_1 - a + b'$, $c'' = c + c'$, and $d'' = n_0 + d' - c$. The posterior density of the exposure odds ratio $\psi_e = \frac{p_1(1-p_0)}{p_0(1-p_1)}$ (Marshall 1988) is

$$p_{\psi_e}(\psi|T'') = \begin{cases} \dfrac{\psi^{a''-1}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\psi)^{a''+b''}}\,dy, \\ \qquad 0 < \psi < 1 \\ \dfrac{\psi^{-(c''+1)}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+\frac{y}{\psi})^{c''+d''}}\,dy, \\ \qquad \psi \geq 1, \end{cases} \tag{6}$$

where $C = \mathbf{B}(a'', b'')\mathbf{B}(c'', d'')$, a product of beta functions.

In Appendix A we show that the marginal posterior density, given by (6), is unimodal. This property is important when calculating HPD intervals using the algorithm of Chen and Shao (1999) (see Sec. 3.3), in preference to the less efficient but more accurate algorithms of Tanner (1993) and Hyndman (1996).

The marginal joint mass function of the numbers exposed in our two independent samples of cases and controls is easily seen to be

$$p_T(T|n_1, n_0, T') = \binom{n_1}{a} \frac{\mathbf{B}(a'', b'')}{\mathbf{B}(a', b')} \times \binom{n_0}{c} \frac{\mathbf{B}(c'', d'')}{\mathbf{B}(c', d')}, \tag{7}$$

where $T \in \mathcal{T} = \{(a, n_1 - a, c, n_0 - c), a = 0, \ldots, n_1$ and $c = 0, \ldots, n_0\}$ and $T' = (a', b', c', d')$.

Theorem B.1 in Appendix B implies that $\frac{a}{n_1} \to^d p_1$ and $\frac{c}{n_0} \to^d p_0$ as n_1 and n_0 approach ∞, where $\to^d$ denotes convergence in distribution. This, together with the independence between a and c, implies that $(\frac{a}{n_1}, \frac{c}{n_0}) \to^d (p_1, p_0)$. This asymptotic property of the marginal distribution is essential to the derivation of Theorem 1 in the next section.

3.2 Approximate ALC_k Sample Size Formulas for the Odds Ratio

The control-to-case ratio can either be based on practical considerations or chosen to minimize the total sample size $n_1 + gn_1$. We use the notation $n_0(g)$ and $n_1(g)$ to emphasize the dependence of n_1 and n_0 on g. When g is known, the ALC_k sample size problem for a case-control design becomes

$$\text{minimize} \quad N_{\psi_e}(g) = n_1(g) + n_0(g) = (g+1)n_1(g), \tag{8}$$

such that

$$\begin{aligned} &\text{alc}_k\big(n_1(g), n_0(g), T'\big) \\ &\quad = \left(\int_{\mathcal{T}_g} \big\{l^\star_{1-\alpha}\big(T|n_1(g), n_0(g), T'\big)\big\}^k \right. \\ &\qquad \left. \times\, p_T\big(T|n_1(g), n_0(g), T'\big)\,dx\right)^{1/k} \leq l, \end{aligned} \tag{9}$$

where $l^\star_{1-\alpha}(T|n_1(g), n_0(g), T')$ refers to the length of the HPD intervals of fixed coverage $1 - \alpha$ under $p_{\psi_e}(\psi|T'' = T + T')$ and $T \in \mathcal{T}_g = \{(a, n_1(g) - a, c, n_0(g) - c), a = 0, \ldots, n_1(g)$ and $c = 0, \ldots, n_0(g)\}$.

When g is not specified in advance, we need to minimize (8) over both g and $n_1(g)$. One way to proceed is to create a grid of values $g_j, j = 1, \ldots, J$, for g, and for each g_j, solve the sample size problem as if $g = g_j$ were known. The overall minimum value is then chosen as the optimal sample size.

We initially regard g as fixed. The HPD length $l^\star_{1-\alpha}(T|n_1(g), n_0(g), T')$ in (8) does not have a closed-form expression, and even a first-order approximation to the credible interval length requires intensive computation. We therefore proceed as follows.

The expression for $\text{alc}_k(n_1(g), n_0(g), T')$ given in (9) is $\{\mathbf{E}_T[l^\star_{1-\alpha}(T|n_1(g), n_0(g), T')]^k\}^{1/k}$, where $\mathbf{E}_T$ denotes expectation with respect to $p_T(T|n_1(g), n_0(g), T')$. We establish an asymptotic expression for $\{\mathbf{E}_T[\widehat{l^\star}_{1-\alpha}(T|n_1(g), n_0(g), T')]^k\}^{1/k}$ where $\widehat{l^\star}_{1-\alpha}(T|n_1(g), n_0(g), T') = 2z_{1-\alpha/2}\sqrt{\text{var}(\psi_e|T'')}$ and $z_{1-\alpha/2}$ is the usual $100(1 - \alpha/2)$ percentile of the normal distribution, and use this expression as an approximation to the criterion function ALC_k. The first-order approximation to $\widehat{l^\star}_{1-\alpha}(T|n_1(g), n_0(g), T')$ is valid when $b', c > 2$.

Theorem 1. Let $k_{\psi_e} = [3(k+1)/2]$. For $a', d' > 0$ and $b', c' > k_{\psi_e}$,

$$\begin{aligned} &\lim_{n_1(g)\to\infty} \frac{\sqrt{n_1(g)}\{\mathbf{E}_T[\widehat{l^\star}_{1-\alpha}(T|n_1(g), n_0(g), T')]^k\}^{1/k}}{2z_{1-\alpha/2}} \\ &\quad = \left\{\int_0^1 \int_0^1 \left[\frac{x(1-x)}{g} + y(1-y)\right]^{k/2} \right. \\ &\qquad \times \frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{\mathbf{B}(a', b')} \\ &\qquad \left. \times \frac{y^{c'-3k/2-1}(1-y)^{d'+k/2-1}}{\mathbf{B}(c', d')}\,dx\,dy\right\}^{1/k}. \end{aligned} \tag{10}$$

For the proof see Appendix C.

The following corollary yields a closed-form approximation to the sample size based on the ALC_k criterion, for fixed g.

Corollary 1. Under the same conditions as Theorem 1, the approximate ALC_k sample size is given by

$$\begin{aligned} N_{\psi_e}(g) &= (g+1)\frac{4z^2_{1-\alpha/2}}{l^2} \\ &\quad \times \left\{\int_0^1 \int_0^1 \left[\frac{x(1-x)}{g} + y(1-y)\right]^{k/2}\right. \end{aligned}$$

$$\times \frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{\mathbf{B}(a', b')} \times \frac{y^{c'-3k/2-1}(1-y)^{d'+k/2-1}}{\mathbf{B}(c', d')}\, dx\, dy \Bigg\}^{2/k} - N_0, \tag{11}$$

where $a', d' > 0$; $b', c' > k_{\psi_e}$; and $N_0 = a' + b' + c' + d'$ is the prior sample size.

Proof. Using the right side of (10) as an approximation to the expression in the square brackets on the left side, and solving for $n_1(g)$ we get (11).

Finally, $N_{\psi_e}(g)$ may be examined as a function of g, to ascertain the optimal $(N_{\psi_e}(g), g)$ combination.

Example 1. Let $k = 1$ and $a' = b' = c' = d' = 3$, which represent the smallest allowable values in Theorem 1 for this choice of k. Suppose that we desire an expected HPD length of $l = 2$, with a coverage of $1 - \alpha = .95$. Figure 1 displays a plot of $N_{\psi_e}(g)$ versus g, where $N_{\psi_e}(g)$ is obtained by (11) with Monte Carlo integration. The optimal g is clearly close to 1, which is not surprising given the symmetry of the prior distribution. The optimal ratios, $g_{\text{opt}} = .975, .99, 1.0, 1.015, 1.025$, all yield the same (to the nearest integer) minimum, $N_{\psi_e}(g_{\text{opt}}) = 472$.

Example 2. Let $k = 2$. It is easy to show, by differentiating $N_{\psi_e}(g)$, given by expression (11), with respect to g, that the optimal ratio is $g = \sqrt{\eta_{\psi_e}}$ when $b', c' > 4.5$ and where

$$\eta_{\psi_e} = \sqrt{\frac{\mathbf{B}(a'+2, b'-2)\mathbf{B}(c'-3, d'+1)}{\mathbf{B}(a'+1, b'-3)\mathbf{B}(c'-2, d'+2)}}.$$

In the context of frequentist hypothesis testing, Gail, Williams, Byar, and Brown (1976) demonstrated that the optimal ratio g follows a "square root rule," that is, $g = \sqrt{\eta}$, where $\eta = \frac{p_1(1-p_1)}{p_0(1-p_0)}$. Consequently, from a Bayesian viewpoint, there is also a square root rule when $k = 2$.

Example 3. When $k = \infty$, the criterion function corresponding to the WOC = ALC_∞ does not converge to 0 as the sample size derived from Theorem 2 increases to infinity. As often occurs with the WOC (see Joseph and Bélisle 1997), even as the sample size increases, there is always at least one dataset for which the length exceeds the target value. The MWOC is preferred in these cases.

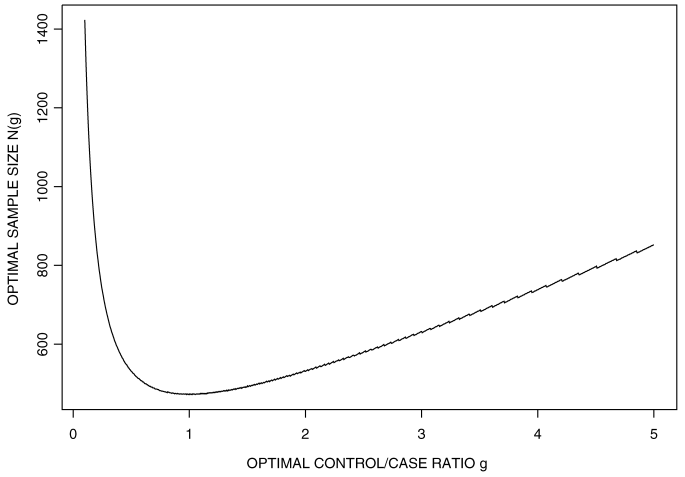

Figure 1. Graph of $N_{\psi_e}(g)$ Against g for the Odds Ratio With $(a', b', c', d', 1-\alpha, l) = (3.0, 3.0, 3.0, 3.0, .95, 2.0)$. The optimal ratios are $g_{opt} \in [.975, 1.025]$, all yielding the same sample size, $N_{\psi_e}(g_{opt}) = 472$.

3.3 Extensions to Other Criteria

So far we have presented a closed-form formula for ALC_k ($k < \infty$) sample sizes when $b', c' > k_{\psi_e}$. There are several other scenarios to consider:

a. Under the ALC_k when b' or $c' < k_{\psi_e}$, the foregoing methods break down.

b. For the ACC_k, one might consider mimicking the approach for the ALC_k instead, minimizing the sum $N_{\psi_e}(g) = n_1(g) + n_0(g) = (g+1)n_1(g)$ subject to the constraint

$$\text{acc}_k\big(n_1(g), n_0(g), T'\big) = \mathbf{E}_T\big[\big\{\alpha_l^\star\big(T|n_1(g), n_0(g), T'\big)\big\}^k\big]^{1/k} \geq 1 - \alpha, \tag{12}$$

where $\alpha_l^\star(T|n_1(g), n_0(g), T')$ is the posterior coverage of an HPD interval of fixed length l under $p_{\psi_e}(\psi|T'' = T + T')$. Unfortunately, unlike the ALC_k, empirical evidence suggests that a first-order approximation does not provide accurate sample sizes.

c. There appear to be no closed-form sample size expressions for the remaining criteria, the MWOC, MLC, and MCC.

For these cases, we turn to Monte Carlo methods.

3.4 Sample Size Determination via Crude Monte Carlo Simulation

We begin by indicating briefly how a simple Monte Carlo approach may be used to derive sample sizes for estimating ψ_e, beginning with the ALC_k. (See Joseph et al. 1995 and Wang and Gelfand 2002 for more details about algorithms that are similar to those discussed here.) Regression-based approaches that are generally more accurate and efficient are described in Section 3.5.

Fix the control-to-case ratio g, the prior parameters, $T' = (a', b', c', d')$, the posterior coverage $1 - \alpha$, and an initial value for $n_1(g)$ and hence, automatically, for $n_0(g)$. The following algorithm cycles through various values of $n_1(g)$:

1. Simulate $p_1^i \sim \text{Be}(a', b')$ and $p_0^i \sim \text{Be}(c', d')$, $i = 1, \ldots, m$.
2. For each pair (p_1^i, p_0^i), simulate two independent observations, $a_i \sim \text{Bin}(n_1(g), p_1^i)$ and $c_i \sim \text{Bin}(gn_1(g), p_0^i)$.
3. For each i, simulate $p_1^j \sim \text{Be}(a_i + a', n_1(g) - a_i + b')$ and $p_0^j \sim \text{Be}(c_i + c', gn_1(g) - c_i + d')$ and set $\psi_j = \frac{p_1^j(1-p_0^j)}{p_0^j(1-p_1^j)}$, $j = 1, \ldots, M$. Estimate the HPD length by, for example, $l_i = \min_{1 \leq j \leq M-[(1-\alpha)M]}(\psi_{(j+[(1-\alpha)M])} - \psi_{(j)})$, according to the approach of Chen and Shao (1999).
4. Compute $\text{alc}_k(n_1(g), n_0(g), T') \approx (\frac{1}{m}\sum_{i=1}^m l_i^k)^{1/k}$; see (9).
5. Cycle through the foregoing steps using a bisectional search strategy until the optimal $n_1(g)$ is attained.

For the ACC_k, one may similarly use Monte Carlo integration to approximate the expression on the right side of (12). Note that we can define

$$\alpha_l^\star = \max_{1\le j\le M} \frac{\#\{1 \le k \le M : \psi_j \le \psi_k \le \psi_j + l\}}{M}.$$

When g is unknown, the foregoing algorithms can be applied over a grid of preselected values for g to estimate the optimal case-control ratio and resulting sample size.

Almost identical algorithms can be applied for the MCC and MLC. To reduce Monte Carlo errors, one could average over a fixed number of estimated sample sizes. Our experience in case-control studies is that the crude Monte Carlo approach has stable trial-to-trial variability of sample size estimates and works well for the coverage criteria ACC_k, MCC, and MLC, but has several limitations when dealing with the length criterion ALC_k. Even the most efficient algorithm for HPD intervals of Chen and Shao (1999) is often too slow in practice. As is shown in Figure 2(a), a prohibitively large number ($m = M > 40{,}000$) of independent runs may be required to reduce trial-to-trial sample size variability to a reasonable magnitude. The bisectional search is a further computational burden, because often numerous steps are required. One difference between coverage-based versus length-based criteria is that coverages are always bounded by 0 and 1, whereas lengths, particularly for the odds ratio, are unbounded and highly variable. An alternative is to adapt a version of the regression-based approach to Bayesian design developed by Müller and Parmigiani (1995) and Müller (1999) to the present situation. Besides producing a stable algorithm for the ALC_k, it also provides a useful method for the other criteria.

3.5 A Regression-Based Approach to Sample Size Determination

We first consider the ALC_k. Again, fix g. When $b', c' > k_{\psi_e}$, the forms of (10) and (11) suggest fitting a regression of the type

$$\mathrm{alc}_k\big(n_1(g), n_0(g), T'\big) = e_1 \frac{1}{n_1^{1/2}} \tag{13}$$

or

$$\frac{1}{\mathrm{alc}_k^2(n_1(g), n_0(g), T')} = e_1 + e_2 n_1, \tag{14}$$

to J Monte Carlo samples $\widetilde{\mathrm{alc}}_k(n_{1j}(g), n_{0j}(g), T'), j = 1, \ldots, J$, where e_1 and e_2 are unknown regression coefficients. These may be estimated using, for example, least squares, to obtain $\hat{e}_1$ and $\hat{e}_2$. We prefer using (14) over (13) because the former better visualizes the linear relationship, as seen in Figure 2(b).

We obtain the sample size by solving the equation $\hat{e}_1 + \hat{e}_2 n_1 = \frac{1}{l^2}$ to get $\hat{n}_1 = \frac{1-\hat{e}_1 l^2}{\hat{e}_2 l^2}$. Even though Theorem 2 was established under the condition that $b', c' > k_{\psi_e}$, we have found

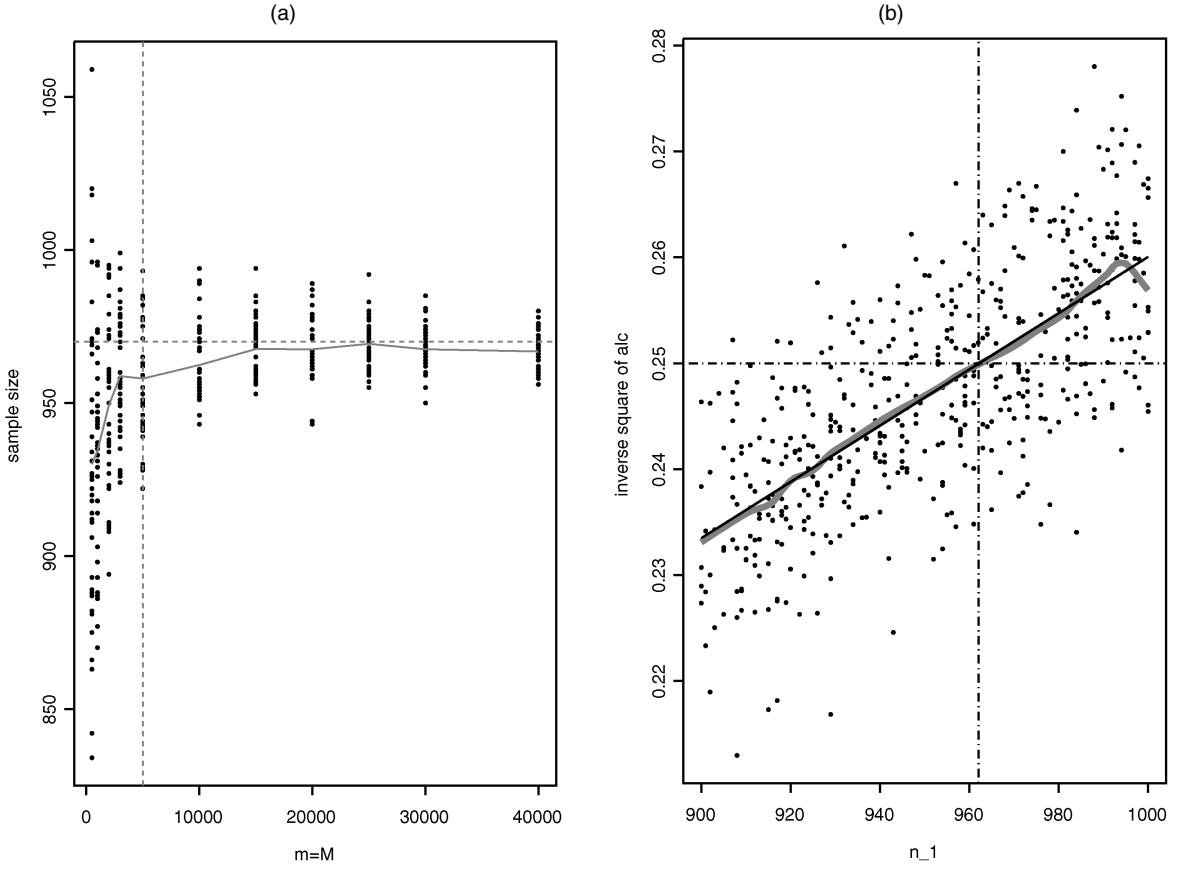

Figure 2. Monte Carlo Graphs. (a) The crude Monte Carlo sample size estimates, $n_1(2)$, plotted against $500 \le m = M \le 40{,}000$ for $(a', b', c', d') = (5, 5, 10, 40)$ and $1 - \alpha = .95$. The horizontal line corresponds to $y = 970$ and vertical line corresponds to $m = M = 5{,}000$. (b) The Monte Carlo pairs $(n_1(2), 1/\widetilde{alc}^2(n_1(2), 5, 5, 10, 40, .95))$. The horizontal line corresponds to $y = .25$ (or $l = 2$), and vertical line corresponds to $n_1 = 962$. (■ linear regression; ■ supersmooth.)

empirically that a linear regression equation fits even when $k < b'$, $c' \le k_{\psi_e}$. The convergence rate is $1/n_1^{\lambda(a',b',c',d')}$ instead of the rate of $1/n_1^{1/2}$ found under (13). The constant $\lambda(a', b', c', d')$ can be recovered by fitting, for example, a regression equation of the form

$$\log\bigl(\text{alc}_k\bigl(n_{1j}(g), n_{0j}(g), T'\bigr)\bigr) = \mu - \lambda \log(n_1). \qquad (15)$$

This leads to a third regression approach, with $\hat{\mu}$ and $\hat{\lambda}$ as least squares estimates and a sample size of $\hat{n}_1 = \exp(\frac{\hat{\mu} - \log(l)}{\hat{\lambda}})$.

Example 4. To illustrate, let the prior information be $T' = \{5, 5, 10, 40\}$ and let $1 - \alpha = .95$, $l = 2.0$, $k = 1$, and $g = 2$. All three regression models give sample sizes very close to $\hat{n}_1 \approx 961$. The estimate of λ is .49, implying a convergence rate of roughly $1/n_1^{1/2}$, as predicted by Theorem 2. In comparison, the closed form given by (11) gives $n_1 = 981$. We used $m = M = 5{,}000$.

Although sample sizes from the different methods were similar, the regression-based approaches were more efficient than the crude Monte Carlo approaches. For the latter approach, we averaged 10 Monte Carlo estimates, each with $m = M = 5{,}000$. The ALC regression-based approach used $J = 100$ uniformly chosen values, $900 \le n_{1j} \le 1{,}000$, and again $m = M = 5{,}000$. The trial-to-trial standard deviation about n_1 was approximately 3, compared with about 20 using the crude Monte Carlo estimates.

In summary, we prefer a regression approach for the ALC_k, which reduces the "noise" inherent in Monte Carlo methods through the use of parametric curves to determine sample sizes. Our approximate sample size formula provides good starting values for the crude Monte Carlo estimates. We carried out a small simulation study that demonstrated that the regression-based approach also works well for the MLC and MWOC. Although similar methods can be developed for the ACC and MCC, the optimal form of the regression equation is not obvious. We therefore rely on fitting a robust smoother, as originally suggested by Müller and Parmigiani (1995) and Müller (1999), to the estimated average and median coverage criterion functions, which we find performs particularly well for the ACC and MCC. The crude Monte Carlo approach seems also to perform well for these two coverage criteria, sometimes even for m and M as small as 500.

3.6 An Illustrative Example

We use an illustrative example contrasting two frequentist and our Bayesian sample size methods for case-control studies. The first frequentist approach, from O'Neill (1984), estimates the sample size as

$$N_{\psi_e}(g) = (g+1)\frac{z_{1-\alpha/2}^2}{\hat{p}_0(1-\hat{p}_0)}\frac{\left\{\frac{(1-\hat{p}_0+\hat{p}_0\widehat{\psi}_e)^2}{\widehat{\psi}_e} + \frac{1}{g}\right\}}{\left\{\text{arcsinh}(\frac{l}{2\widehat{\psi}_e})\right\}^2}, \qquad (16)$$

where l is the desired length of a $100(1-\alpha)\%$ confidence interval around the odds ratio $\widehat{\psi}_e$ and $\hat{p}_0$ is the assumed proportion exposed in the control group.

The sample size given by (16) does not take into account the stochastic nature of confidence intervals or the uncertainty in the values of $\widehat{\psi}_e$ or $\widehat{p}_0$. To address the former shortcoming, Satten and Kupper (1990) proposed choosing N such that the anticipated reported confidence interval has length no more than l with probability $1 - \gamma$. In the case-control context, this requires that one find the minimal $n_1(g)$ such that

$$\Pr\Biggl[\frac{a(n_0(g)-c)}{(n_1(g)-a)c} \times \sinh\left(z_{1-\alpha/2}\sqrt{\text{var}\left(\frac{1}{a} + \frac{1}{n_1(g)-a} + \frac{1}{c} + \frac{1}{n_0(g)-c}\right)}\right) \le \frac{l}{2}\Biggr] \ge 1-\gamma, \qquad (17)$$

where a and c represent the random variables whose outcomes are $0 \le a \le n_1(g)$ and $0 \le c \le n_0(g)$ (Table 1). The left side of (17) may be estimated by means of Monte Carlo simulations, and a bisectional search may be used to find the sample size $N = (g+1)n_1(g)$.

Multiple Sclerosis and the Epstein–Barr Virus. There may be an association between Epstein–Barr virus (EBV) infection and subsequent development of multiple sclerosis (MS) (Marrie et al. 2000). EBV causes mononucleosis in a small proportion of those who are infected. In a case-control study of this hypothesis, n_1 subjects with MS (the cases) would be compared with a group of n_0 subjects (suitably chosen) without MS (the controls). Antibody titers would be used to ascertain which of the cases and controls had previously been exposed to EBV. Let p_1 and p_0 represent the proportion of subjects who were exposed to EBV among the cases and the controls. Suppose that a pilot study revealed that out of 7 subjects with MS and 16 without MS, the numbers of exposed to EBV were 3 and 4.

From a Bayesian perspective, one may ask the following question: Using the prior information gained from the pilot study, what should the sample sizes of the main study be to ensure that a 95% posterior credible interval for the odds ratio $\psi_e = \frac{p_1(1-p_0)}{(1-p_1)p_0}$ would have a length no larger than, say, 3? From a frequentist perspective, based on (16), one would seek the sample size that gives a 95% confidence interval for ψ_e of length 3, using data from the pilot study to estimate p_0, q_0, and ψ_e in the expression in the right side of (16). Alternatively, one could use (17) with the specification of $l = 3$ ($l/2 = 1.5$), $1 - \alpha = .95$, and $1 - \gamma = .9$, again using the pilot study to simulate a and c.

For the Bayesian analysis, we set the values for the prior parameters to be $(a', b', c', d') = (3, 4, 4, 12)$, as suggested by the pilot study. Similarly, for both frequentist approaches we used $\widehat{p}_0 = .25$ and $\widehat{p}_1 = 3/7$.

Results. The optimal frequentist values of g were 1.14 based on (16) and 1.24 based on (17). Table 2 presents sample sizes based on (16) and (17), including, for comparison, sample sizes for $g = 1, 2$. The sample sizes from (16) were, as expected, less than those from (17), because the stochastic nature of the data were ignored in (16). The optimal sample size did not differ much between the optimal ratio and a simple 1 : 1 ratio of controls to cases.

Table 2. Comparative Table of Frequentist Sample Sizes for $g = 1.14$ and 2, $1-\gamma = .90$, $1-\alpha = .95$, and $l = 3.0$

g	n_1	n_0	N
From O'Neill, (16)			
2.00	193	386	579
1.24	240	297	537
1.14	251	286	537
1.00	269	269	538
From Satten and Kupper, (17)			
2.00	292	584	876
1.24	365	453	818
1.14	383	437	820
1.00	413	413	826

Table 4. Comparative MWOC Table of Bayesian Sample Sizes for $g = 1.0$, 2.0 and $1-\gamma = .50, .75, .85, .90$; $1-\alpha = .95$, $l = 3.0$, and $(a', b', c', d') = (3, 4, 4, 12)$

$1-\gamma$	n_1	n_0	N
$g = 1.0$			
.50	1,409	1,409	2,818
.75	4,635	4,635	9,270
.85	9,241	9,241	18,482
.90	12,750	12,750	25,500
$g = 2.0$			
.50	943	1,886	2,829
.75	2,987	5,974	8,961
.85	6,055	12,110	18,165
.90	8,221	16,442	24,663

The optimal Bayesian value of g was near 1.14, regardless of the criterion used. Table 3 presents Bayesian sample sizes for the ALC from both (11) and from Monte Carlo methods, as well as sample sizes from the ALC_2, ACC, MLC, and PGT-(ii). Table 4 presents results from the MWOC for a variety of coverages $(1-\gamma)$. The MLC gave by far the smallest sample sizes, below $n_0 + n_1 = 200$ regardless of the value of g. In contrast, much larger values, in the range 600–700, were seen from the ALC. The ACC provided even higher values than the ALC. At first glance, it may be surprising that the MLC sizes were considerably smaller than the ACC or ALC sizes, which average over the predictive distribution of the data. Clearly, the distribution of lengths and coverages across possible datasets were highly skewed, with a least half of the datasets leading to relatively narrow HPD intervals compared with those intervals from other datasets. Further, the ACC led to higher sample sizes compared with the ALC, because ensuring an average coverage of .95 near the boundary of 1 is more difficult that averaging lengths of 3, which are not near to an endpoint of the range of possible lengths (see Joseph et al. 1995 for further discussion). The ALC_2 and PGT-(ii) suggested sample sizes >3,000, demonstrating that variance is a more difficult parameter to control compared with lengths or coverages. Even so, all of these criteria provided the desired accuracy only on average (or in the median), so it is not surprising that the MWOC suggested even larger sample sizes, as displayed in Table 4.

Table 3. Comparative Table of Bayesian Sample Sizes for Various g; $1-\alpha = .95$, $l = 3.0$, and $(a', b', c', d') = (3, 4, 4, 12)$

g	n_1	n_0	N
ALC from (11)			
2.00	242	482	724
1.24	312	380	692
1.14	327	365	692
1.00	355	346	701
ALC			
2.00	226	452	678
1.24	282	352	634
1.14	297	339	636
1.00	323	323	646
ACC			
2.00	503	1,006	1,509
1.24	643	798	1,441
1.14	676	771	1,447
1.00	736	736	1,472
g	n_1	n_0	$N(g)$
MLC			
2.00	64	128	192
1.24	79	98	177
1.14	82	94	176
1.00	88	88	176
ALC_2			
2.00	1,150	2,300	3,450
1.24	1,512	1,875	3,387
1.14	1,599	1,823	3,422
1.00	1,736	1,736	3,472
PGT-(ii)			
2.00	1,328	2,656	3,984
1.24	1,743	2,162	3,905
1.14	1,838	2,096	3,934
1.00	2,004	2,004	4,008

More generally, in this example, for a wide range of g, we found the following ordering of sample sizes:

$$\text{MCC} \approx \text{MLC} < \text{O'Neill} < \text{ALC} < \text{Satten and Kupper} < \text{ACC}.$$

Although this ordering may not hold for a different set of prior parameters, our sample size formulas for the ALC and the Monte Carlo regression-based approach seem to suggest that the orderings obtained using the length criteria ALC, MLC, MWOC, and O'Neill are approximately independent of α and l.

Overall, we recommend that when designing a study, sample sizes should be calculated using a variety of criteria, and full use should be made of the prior information available. In this example it would be quite risky to rely on the MLC sample size, because HPD widths much larger than desired would arise with high probability. Similarly, sample sizes using (16) are likely insufficient, because the stochastic nature of the data is ignored. At the other extreme, there is no strong reason to prefer the ALC_2 [or PGT-(ii)] over ALC, with the latter depending on a more natural metric. Different prior information could lead to different orderings of the sizes provided by the various criteria, so that the foregoing conclusions will not necessarily hold in other examples.

The very wide range in sample sizes here is explained in large part by the small amount of prior information. This in turn allows for very wide ranges of p_0 and p_1 values, meaning that the odds ratios can vary considerably from sample to sample. Next we discuss the influence of varying prior information on sample size determination.

3.7 Sensitivity Analysis

We carried out two different sensitivity analyses. First, we examined the effect of different prior specifications from our example in Section 3.6. Second, and unrelated to our example,

we examined the effect of small changes in prior parameter values.

We considered two alternative specifications for our example in Section 3.6: (1.5, 2, 2, 6), which provides one-half of the data-equivalent information compared with our original prior values of (3, 4, 4, 12) and (6, 8, 8, 24), which provides twice the data-equivalent information. We again examined the four criteria ALC, ACC, MLC, and MCC, with $1-\alpha = .95$ and $l = 3$.

The sample sizes from each choice of prior specification are provided in Table 5. Only slight sample size reductions were realized for $g = 1$ compared to $g = 2$. Comparing the sample sizes arising from the different prior distributions, it is clear that the amount of prior information exerted a significant influence, although we also note that the prior mean odds ratio also changed as the prior information changed, also contributing to the effect. More precise prior information led to smaller sample sizes. We also note that the MLC and MCC led to similar sample sizes, and that the ACC led to the largest sample size.

The ALC sample sizes obtained under the quite diffuse specification (1.5, 2, 2, 6) were unstable. For example, with $m = M = 10{,}000$, 15,000, 20,000 and $g = 1$, (13) led to a sample size of $n_1 =$ 5,071, 5,445, 4,947. Equation (14) led to $n_1 =$ 4,521, 4,722, 4,648, and (15) let to $n_1 =$ 4,784, 5,054, 4,824. The finding that the size formula did not work well for the prior table (1.5, 2, 2, 6) was expected, because this prior specification does not satisfy the conditions required by these equations.

For the second sensitivity analysis, we considered the ALC, ACC, and MWOC, with $1-\alpha-.95$, $l = 1$, $g = 1$, and $\gamma = .5$, .75, and .85. The prior specifications under comparison were (50, 50, 50, 50) and (49, 49, 51, 51). These somewhat informative priors led to the same prior sample size of $N_0 = 100$. The corresponding total sample sizes [ALC, ACC, MWOC(.5), MWOC(.75), MWOC(.85)] were (428, 481, 850, 1,150, 1,366) for the first prior specification and (428, 480, 851, 1,150, 1,372) for the second prior specification. Therefore, our sample sizes were robust to small changes in the prior specification. In addition, the ALC and MWOC sample sizes were similar regardless of the method used to calculate them.

4. MORE GENERAL CASE–CONTROL SETTINGS

We now extend the prototypic case-control design considered earlier in various directions. We first consider cross-sectional designs, and then show how our methods can be applied to gene–environment studies.

4.1 Cross-Sectional Studies

In a cross-sectional study, disease and exposure outcomes are measured simultaneously. Let $p_{11} = \Pr(D = 1, E = 1)$, $p_{10} = \Pr(D = 1, E = 0)$, $p_{01} = \Pr(D = 0, E = 1)$, and $p_{00} = \Pr(D = 0, E = 0) = 1 - p_{11} - p_{10} - p_{01}$. The odds ratio is given by $\psi = \frac{p_{11}p_{00}}{p_{10}p_{01}}$.

The usual Bayesian approach in a cross-sectional study is to regard the likelihood realizations $T = (a, b, c, d)$, for fixed $N = a + b + c + d$ (see Table 1) as being governed by a multinomial distribution, $\text{Mult}(N; p_{11}, p_{10}, p_{01}, p_{00})$, and the probabilities $(p_{11}, p_{10}, p_{01}, p_{00})$ as Dirichlet, $\text{Dir}(a', b', c', d')$. Under this Dirichlet-multinomial specification, the posterior density of ψ is the same as the posterior density of ψ_e given earlier by (6) (Latorre 1982, 1984). The consequent unimodality of the posterior density, as we have noted earlier, facilitates implementation of the algorithm of Chen and Shao (1999).

We consider the problem of determining the required sample size $N = a + b + c + d$. Let

$$\mathbf{B}(a_1, a_2, \ldots, a_j) = \frac{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_j)}{\Gamma(a_1 + a_2 + \cdots + a_j)}$$

be the multivariate beta function and let $a'' = a + a', b'' = b + b', c'' = c + c'$, and $d'' = d + d'$. The marginal probability function of $T = (a, b, c, d)$ is

$$p_G(T|N, T') = \binom{N}{a\ b\ c\ d} \frac{\mathbf{B}(a'', b'', c'', d'')}{\mathbf{B}(a', b', c', d')}. \tag{18}$$

Theorem B.1 implies that $(\frac{a}{N}, \frac{b}{N}, \frac{c}{N}, \frac{d}{N}) \to^d (p_{11}, p_{10}, p_{01}, p_{00})$ as $N \to \infty$, a result essential to the derivation of the sample size formula given by (20).

If we reconsider the parameterization (n_1, a, c) where the random variable n_1 is $a + b$, then a straightforward change of variable in (18) leads to

$$\begin{aligned} &p_G(n_1, a, c|N, T') \\ &\quad = \binom{N}{n_1} \frac{\mathbf{B}(n_1 + a' + b', n_0 + c' + d')}{\mathbf{B}(a' + b', c' + d')} p_T(T|n_1, n_0, T'), \\ &\quad n_1 = 0, 1, \ldots, N, a = 0, 1, \ldots, n_1, \text{ and} \\ &\quad c = 0, 1, \ldots, n_0, n_0 = N - n_1, \end{aligned} \tag{19}$$

Table 5. Comparative Table of Bayesian Sample Sizes for (a′, b′, c′, d′) = (6, 8, 8, 24), (3, 4, 4, 12), (1.5, 2, 2, 6), $1-\alpha = .95$, and $l = 1.0$

		(6, 8, 8, 24)			(3, 4, 4, 12)			(1.5, 2, 2, 6)		
Criterion	g	n_1	n_0	N	n_1	n_0	N	n_1	n_0	N
ALC from (11)	2.0	97	190	287	242	482	724	3,791	7,582	11,373
	1.0	144	126	270	355	346	701	5,707	5,703	11,410
ALC	2.0	89	178	267	226	452	678	3,034	6,068	9,102
	1.0	124	124	248	323	323	646	4,521	4,521	9,042
ACC	2.0	159	318	477	503	1,006	1,509	3,645	7,290	10,935
	1.0	227	227	454	736	736	1,472	5,422	5,422	10,844
MLC	2.0	51	102	153	64	128	192	80	160	240
	1.0	69	69	138	88	88	176	111	111	222
MCC	2.0	51	102	153	62	124	186	77	154	231
	1.0	68	68	136	88	88	176	111	111	222

where $p_T(T|n_1, n_0, T')$ is as defined in (7). This form allows us to once again use Monte Carlo simulations to find the optimal N. The details are omitted here (see M'Lan 2002).

The main difference between the case-control setting and the cross-sectional design is that the control-to-case ratio is no longer fixed, but rather is random with $\frac{N}{g+1} \sim \text{Mult}(N; \pi)$, where $\pi \sim \text{Be}(a' + b', c' + d')$. Using the same methods as in Section 3.2, it follows that there is a closed-form expression for the approximate sample size for the ALC_k,

$$N = \frac{4z_{1-\alpha/2}^2}{l^2} \times \left\{ \int_0^1 \left\{ \int_0^1 \int_0^1 \left[\frac{1}{1-\pi} x(1-x) + \frac{1}{\pi} y(1-y) \right]^{k/2} \times \frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{\mathbf{B}(a', b')} \times \frac{y^{c'-3k/2}(1-y)^{d'+k/2-1}}{\mathbf{B}(c', d')} \, dx \, dy \right\} \times \frac{\pi^{a'+b'-1}(1-\pi)^{c'+d'-1}}{\mathbf{B}(a'+b', c'+d')} \, d\pi \right\}^{2/k} - N_0, \quad (20)$$

where $a', d' > 0$ and $b', c' > k_\psi = [3(k+1)/2]$, and $N_0 = a' + b' + c' + d'$ is the prior sample size. [For the derivation of (20), see M'Lan 2002.] Whereas the foregoing closed-form expression for the sample size is available for the ALC, sample sizes from all other criteria can be derived via the Monte Carlo methods similar to those outlined in Sections 3.4 and 3.5.

4.2 Gene–Environment Studies

The importance of gene–environment studies has coincided with growing recognition that many human diseases are the result of a joint effect of genes and the environment. In this section we extend and adapt the results of the preceding sections to the design of case-control studies whose purpose is to examine gene–environment effects defined in a precise way. We concentrate on 2×4 case-control designs as well as gene–environment case-only designs.

Let G be a dichotomous variable indicating the presence ($G = 1$) or absence ($G = 0$) of an inherited susceptibility genotype. As in any case-control study, we start by collecting n_1 cases and $n_0 = gn_1$ controls. Each subject is classified into one of the four possible pairs of the exposure E and genotype G. Table 6 summarizes the data so collected and the parameters of interest. Cases and controls in the first row of Table 6 form the reference group against which odds ratios will be calculated.

Table 6. Generic 2×4 Table for Gene–Environment Interaction Analysis in Case-Control Settings

Exposure	*Susceptibility genotype*	*Cases*	*Controls*	*Odds ratios*
0	0	A_{00}	B_{00}	1.0
0	1	A_{01}	B_{01}	$\psi_G = \frac{p_{01}q_{00}}{p_{00}q_{01}}$
1	0	A_{10}	B_{10}	$\psi_E = \frac{p_{10}q_{00}}{p_{00}q_{10}}$
1	1	A_{11}	B_{11}	$\psi_{EG} = \frac{p_{11}q_{00}}{p_{00}q_{11}}$
Total		n_1	n_0	

Let $p_{ij} = \Pr(G = i, E = j|D = 1)$ and $q_{ij} = \Pr(G = i, E = j|D = 0)$, $i, j = 0, 1$, be the cell probabilities given disease status. Let ψ_G, ψ_E, and ψ_{GE} denote the odds ratios of disease for $G = 1$ and $E = 0$, $G = 0$ and $E = 1$, and $G = 1$ and $E = 1$, relative to the reference group ($G = 0$, $E = 0$). By definition, we have

$$\psi_G = \frac{\Pr(D=1|G=1,E=0)}{\Pr(D=1|G=0,E=0)} \Big/ \frac{\Pr(D=0|G=1,E=0)}{\Pr(D=0|G=0,E=0)} = \frac{p_{01}q_{00}}{p_{00}q_{01}},$$

$$\psi_E = \frac{\Pr(D=1|G=0,E=1)}{\Pr(D=1|G=0,E=0)} \Big/ \frac{\Pr(D=0|G=0,E=1)}{\Pr(D=0|G=0,E=0)} = \frac{p_{10}q_{00}}{p_{00}q_{10}},$$

and

$$\psi_{GE} = \frac{\Pr(D=1|G=1,E=1)}{\Pr(D=1|G=0,E=0)} \Big/ \frac{\Pr(D=0|G=1,E=1)}{\Pr(D=0|G=0,E=0)} = \frac{p_{11}q_{00}}{p_{00}q_{11}}.$$

One way to measure the influence of gene–environment interactions on disease occurrence is to compute the synergy index, I_{GE}, where

$$I_{GE} = \frac{\psi_{GE}}{\psi_G \psi_E} = \frac{p_{11}p_{00}}{p_{01}p_{10}} \Big/ \frac{q_{11}q_{00}}{q_{01}q_{10}}.$$

We exploit the latter form, which is a ratio of two odds ratios. Piegorsch, Weinberg, and Taylor (1994) showed that I_{GE} is equivalent to the interaction parameter between genotype and environment under a logistic regression model.

The natural prior-likelihood model for a 2×4 gene-environment study assumes that conditional on $(n_1; p_{11}, p_{10}, p_{01}, p_{00})$ and $(n_0; q_{11}, q_{10}, q_{01}, q_{00})$, respectively, $T_1 = (A_{11}, A_{10}, A_{01}, A_{00})$ and $T_0 = (B_{11}, B_{10}, B_{01}, B_{00})$ are independent multinomial random vectors using the notation of Table 6. That is, $T_1 \sim \text{Mult}(n_1; p_{11}, p_{10}, p_{01}, p_{00})$ and $T_0 \sim \text{Mult}(n_0; q_{11}, q_{10}, q_{01}, q_{00})$. Next, we assume that the two multinomial cell probability vectors are independent with different Dirichlet distributions, that is, $(p_{11}, p_{10}, p_{01}, p_{00}) \sim \text{Dir}(a_{00}, a_{10}, a_{01}, a_{00})$, and $(q_{11}, q_{10}, q_{01}, q_{00}) \sim \text{Dir}(b_{00}, b_{10}, b_{01}, b_{00})$.

Latorre (1984) derived lengthy expressions for the posterior distribution of I_{GE} using the sum of four infinite series. Rather than work with these expressions, we again take a Monte Carlo approach, which is computationally more efficient. We showed in Section 4.1 that the posterior distribution of the parameters $\log(\frac{p_{11}p_{00}}{p_{01}p_{10}})$ and $\log(\frac{q_{11}q_{00}}{q_{01}q_{10}})$ are strongly unimodal; therefore, the posterior distribution of their difference is also strongly unimodal, and the unimodality of I_{GE} follows from Theorem A.2. The joint marginal distribution of T_1 and T_0 conditional on (n_1, n_0, T_1', T_0') is $p(T_1, T_0|n_1, n_0, T_1', T_0') = p_G(T_1|n_1, T_1')p_G(T_0|n_0, T_0')$, where $T_1' = (a_{11}, a_{10}, a_{01}, a_{00})$, $T_0' = (b_{11}, b_{10}, b_{01}, b_{00})$, and $p_G(T|N, T')$ is defined in (19).

Once again, a Monte Carlo algorithm, along the lines of those of Section 3.4, yields optimal values for n_1 and g. Here, however, one begins by simulating the vectors $\mathbf{p}^i =$

$(p^i_{11}, p^i_{10}, p^i_{01}, p^i_{00})$ and $\mathbf{q}^i = (q^i_{11}, q^i_{10}, q^i_{01}, q^i_{00})$ from their respective Dirichlet priors.

Case-Only Design. In general, inference about I_{GE} requires information about the controls. Suppose, however, that there is strong theoretical justification or empirical evidence that genotype and exposure occur independently in the population. Under this assumption, and assuming that the disease is rare, Piegorsch et al. (1994) showed that

$$I_{GE} \approx \frac{\Pr(E=1|G=1,D=1)\Pr(E=0|G=0,D=1)}{\Pr(E=0|G=1,D=1)\Pr(E=1|G=0,D=1)} = \frac{p_{11}p_{00}}{p_{01}p_{10}}. \quad (21)$$

Approximation (21) implies that we can estimate I_{GE} using data from cases only, because $\frac{p_{11}p_{00}}{p_{01}p_{10}}$ can be estimated without information on controls. Piegorsch et al. (1994) also showed that estimation of gene–environment interaction with a case-only design, if feasible, offers greater precision than that provided by the traditional approach and avoids the difficult problem of validating the control group.

Inferentially, the case-only design is identical to the cross-sectional sampling design discussed in Section 4.1. Therefore, all Bayesian sample size methods for the cross-ratio parameter ψ apply directly to estimating I_{GE} under this case-only setting.

5. BAYESIAN DECISION–THEORETIC SAMPLE SIZE METHODS

Informally, any sample size problem can be considered a form of a decision problem. In this sense, the preceding work in this article can broadly be considered as decision theoretic. More formally, however, Bayesian decision-theoretic sample size problems should be expressed in the form of maximizing an expected utility function over the set of all possible designs of size $n \geq 0$ and over all terminal decisions $d \in \mathcal{D}$ (Lindley 1997). After observing data $\mathbf{x}_n$, we wish to make a decision $d \in \mathcal{D}$ about our parameter of interest θ. Thus we find the minimal n that maximizes

$$U(n) = \int_{\mathcal{X}_n} \left\{ \max_{d\in\mathcal{D}} \int_\Theta u(n, \mathbf{x}_n, d, \theta) p(\theta|\mathbf{x}_n, n)\, d\theta \right\} p(\mathbf{x}_n|n)\, d\mathbf{x}_n.$$

For the sample size problem, a common form for the utility function $u(n, \mathbf{x}_n, d, \theta)$ is

$$u(n, \mathbf{x}_n, d, \theta) = K\delta(n, \mathbf{x}_n, d, \theta) - Lw(n, \mathbf{x}_n, d) - cn \quad (22)$$

for an interval d, where $\delta(n, \mathbf{x}_n, d, \theta) = 1$ if $\theta \in d$, and $\delta(n, \mathbf{x}_n, d, \theta) = 0$ otherwise, and $w(n, \mathbf{x}_n, d)$ is the width of the interval d. The quantities $K, L > 0$ are positive constants balancing high coverage against low width, and $c \geq 0$ is the common cost associated with observing each subject. When $c = 0$, we have

$$U(n, K, L) = \int_{\mathcal{X}_n} \max_{d\in\mathcal{D}} \left\{ K \int_d p(\theta|\mathbf{x}_n, n)\, d\theta - L \int_d d\theta \right\} p(\mathbf{x}_n|n)\, d\mathbf{x}_n. \quad (23)$$

Our ACC and ALC criteria satisfy

$$\mathrm{acc}(n) = \int_{\mathcal{X}_n} \left\{ \max_{C\in\mathcal{I}(l)} \int_C p(\theta|\mathbf{x}_n, n)\, d\theta \right\} p(\mathbf{x}_n|n)\, d\mathbf{x}_n \geq 1 - \alpha$$

and

$$\mathrm{alc}(n) = \int_{\mathcal{X}_n} \left\{ \min_{C\in\mathcal{I}(1-\alpha)} \int_C d\theta \right\} p(\mathbf{x}_n|n)\, d\mathbf{x}_n \leq l,$$

where $\mathcal{I}(l)$ and $\mathcal{I}(1-\alpha)$ are the sets of all posterior credible intervals of length l and coverage $1-\alpha$. When the parameter space for θ is bounded, it is easily seen that the criterion functions $\mathrm{acc}(n)$ and $\mathrm{alc}(n)$ are limiting cases, as m approaches infinity, of the sequence of functions $U(n, 1, L_m)$ when setting $\mathcal{D} = \mathcal{I}(l)$ and of $U(n, K_m, 1)$ when setting $\mathcal{D} = \mathcal{I}(1-\alpha)$, where the sequences of numbers L_m and K_m converge to 0. These results show that even though our methods do not represent fully decision-theoretic criteria, they are close in form. In addition, the ALC_2 is asymptotically equivalent to the PGT-(ii), which is a fully decision-based criteria associated with a quadratic loss function.

Solving (22) using Monte Carlo methods similar to those described in this article or those described by Müller and Parmigiani (1995) should be feasible. In practice, however, deriving sensible loss functions for case-control studies is a difficult problem except for simple cost functions. (See Joseph and Wolfson 1997 for further discussion of this point.)

Rather than selecting a sample size in advance, one can proceed sequentially. One simple way to do this would be to apply any of the criterion defined in Section 2 but with updated prior distributions, based on the data collected so far. For the case-control setting discussed elsewhere in this article, the updated prior is again a beta distribution by conjugacy. One can then calculate how many further subjects need to be recruited in the case and control groups according to the criterion chosen, and proceed accordingly. Other, more complex sequential sampling schemes can also be considered (see Chen 2000; Berger 1985).

6. CONCLUSION

Considering the central role played by case-control studies in epidemiology, De Santis et al. (2004) provided a timely first presentation of Bayesian sample size methods for such studies. We have extended the methods of De Santis et al. in several ways: We have included six different Bayesian sample size criteria (two of which are new) and based our sample size methods on both HPD and equal-tailed intervals. Our methods, which allow for the simultaneous estimation of the optimal control-to-case ratio, provide a general unified Monte Carlo framework for Bayesian sample size determination in case-control studies. At the same time, in some special cases, we provide a closed-form expression for the sample size. One benefit of this approach is that the expression given by Theorem 1 allows us to describe the rate of convergence of the criterion function $\mathrm{alc}_k(n_1, n_0, T')$ as n_0 $(= gn_1)$ tends to ∞. The extension of our methods beyond 2×2 tables to include 2×4 tables of case-only designs (a mainstay of gene–environment studies) and to cross-sectional studies greatly enhances their usefulness.

Using a Monte Carlo approach, sample size estimates can be made arbitrarily accurate. Software implementing the basic methods along with the various extensions discussed in this article is available from the authors. This software also implements sample size estimation for $\log(\psi_e)$. Approximate ALC_k sample sizes and cost formulas for $\log(\psi_e)$ are obtained by replacing the terms $a' + k/2$, $b' - 3k/2$, $c' - 3k/2$, and $d' + 3k/2$

in (11) and (20) by the terms $a' - k/2, b' - k/2, c' - k/2$, and $d' - k/2$. The software extends the Monte Carlo approaches to sample size to the case where it is known a priori that $\psi_e > 1$ or $\psi_e < 1$. Knowing a priori that $\psi_e > 1$ can reduce the required sample size by up to 80%. This observation again reinforces the advantage of incorporating prior information into the model.

In our software, we also extend our sample size problem to a related cost problem, although the subject is not discussed in any detail in this article. If c_0 represents the unit cost per control and $c_1 = rc_0$ represents the unit cost per case, then the cost problem seeks the minimal $n_1(g)$ for a given g or the minimal $(g, n_1(g))$ when g is unknown that minimizes the objective function $C(n_1) = c_1 n_1 + c_0 n_0 = (g + r)c_0 n_1(g)$. This cost function is simple but still sufficiently realistic for some situations. When g is known, a cost formula is obtained by replacing the term $(g + 1)$ in (11) and (20) by $c_0(g + r)$ and replacing N_0 by $C_0 = c_0(c' + d' + ra' + rb')$.

All prior-likelihood models examined here can be put in the framework of a logistic regression, when the log-odds ratio is the parameter of interest. This opens the door for application of our methods when exposure covariates are continuous, although the prior specification for the exposure parameters must be carefully considered. The implementation often would require using MCMC with automated convergence checks for each sample point in the marginal space.

APPENDIX A: UNIMODALITY OF THE POSTERIOR DENSITY IN (6)

Our proof of the unimodality of the density in (6) requires the notion of strongly unimodal distributions (Dharmadhikari and Joag-Dev 1988) along with the following lemmas.

Lemma A.1 (Dharmadhikariand and Joag-Dev 1988). The set of all strongly unimodal distributions is closed under convolutions. In particular, if X_1 and X_2 are two independent strongly unimodal random variables, then so is $X_2 - X_1$.

Lemma A.2 does not appear to have been stated elsewhere.

Lemma A.2. Let U be a strongly unimodal random variable with an absolutely continuous density f_U. Then $V = \exp(U)$ is unimodal.

Proof. The density of V is easily seen to be $f_V(v) = \frac{f_U(\log(v))}{v}$, and, consequently, $\log(f_V(v)) = \log(f_U(\log(v))) - \log(v)$. The differentiation of $\log(f_V(v))$ with respect to v gives

$$\frac{\partial \log(f_V(v))}{dv} = \frac{1}{v}\left\{\frac{f_U'(\log(v))}{f_U(\log(v))} - 1\right\}, \tag{A.1}$$

where $f_U'(u) = \frac{\partial f_U(u)}{du}$. Because U is strongly unimodal, $\frac{f_U'}{f_U}$ is decreasing. Therefore, the right side of (A.1) can have at most one change in sign. If the sign does change, then the change must be from positive to negative. This shows that $\exp(U)$ is unimodal.

Theorem A.1. Let $p_1 \sim \text{Be}(\alpha_1, \beta_1)$ and $p_0 \sim \text{Be}(\alpha_2, \beta_2)$ be two independent random variables with $\alpha_1, \alpha_0, \beta_1, \beta_0 > 0$. Define $\phi_1 = \log(\frac{p_1}{1-p_1})$, $\phi_0 = \log(\frac{p_0}{1-p_0})$, and $\rho = \frac{p_1(1-p_0)}{p_0(1-p_1)}$. Then (a) ϕ_1 and ϕ_0 are strongly unimodal, (b) $\log(\rho)$ is strongly unimodal (and therefore unimodal), and (c) ρ is unimodal.

Proof. (a) Clearly, the density of ϕ_1, $f_{\phi_1}(\phi) = \frac{e^{\alpha_1 \phi}}{(1+e^{\phi})^{\alpha_1+\beta_1}}$, is strongly unimodal, because

$$\frac{\partial^2 \log(f_{\phi_1}(\phi))}{\partial \phi^2} = -(\alpha_1 + \beta_1)\frac{e^{\phi}}{(1 + e^{\phi})^2}.$$

The same holds true for ϕ_0.

(b) Lemma A.1 ensures that $\log(\rho) = \phi_1 - \phi_0$ is strongly unimodal as the difference of two independent strongly unimodal random variables.

(c) Again, Lemma A.2 implies that $\rho = e^{\log(\rho)}$ is unimodal.

Corollary A.1. The posterior density in (6) is unimodal.

Proof. Under the prior/likelihood model in Section 3.1, the posterior distributions of p_1 and p_0 are $p_1 \sim \text{Be}(a'', b'')$ and $p_0 \sim \text{Be}(c'', d'')$. In addition, p_1 and p_0 are independent. Theorem A.1 clearly ensures that the posterior density for ψ_e is unimodal.

APPENDIX B: ASYMPTOTIC DISTRIBUTION OF THE MARGINAL PREDICTIVE DISTRIBUTION

Theorem B.1, which is a straightforward application of Khintchin's weak law of large numbers, does not seem to have been formulated previously.

Theorem B.1. Let $X_1, \ldots, X_n$ conditional on θ be n iid (possibly multivariate) random variables such that $X_i|\theta \sim f_X(x|\theta), i = 1, \ldots, n$, $\theta \sim f(\theta)$, $E(\|X\||\theta) < \infty$, and let $Z = E(X|\theta)$ except on a set of measure zero with respect to $f(\theta)$. Let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n}{n} \to^P Z. \tag{B.1}$$

Proof. We have

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon\right] \\
&\quad = \lim_{n\to\infty} \int_{\Theta} \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon \,\middle|\, \theta\right] f(\theta)\, d\theta \qquad \text{(B.2)} \\
&\quad = \int_{\Theta} \lim_{n\to\infty} \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon \,\middle|\, \theta\right] f(\theta)\, d\theta \qquad \text{(B.3)} \\
&\quad = 0,
\end{aligned}$$

by first using Lebesgue's dominated convergence theorem to interchange the limit and the integral in expression (B.2) to get (B.3), and then using Khintchin's theorem to evaluate the limit inside the integral.

APPENDIX C: PROOF OF THEOREM 1

Define $Z_{n_1,T'} = (\frac{a}{n_1}, \frac{c}{gn_1}) = (X_{n_1,T'}, Y_{n_1,T'})$ given n_1 and T', and let $\mathcal{F}_{n_1}$ be the set of points (x, y) where the mass function of Z_{n_1} is positive. After some algebraic manipulations and setting $a = n_1 x$, $b = n_1(1 - x)$, $c = gn_1 y$, and $d = gn_1(1 - y)$, we have

$$n_1 \operatorname{var}(\psi_e|T'') = \frac{(x+\frac{a'}{n_1})(1-y+\frac{d'}{gn_1})Q(n_1,x,y,T')}{(1-x+\frac{b'}{n_1})^2(1-x+\frac{b'-2}{n_1})(y+\frac{c'-1}{gn_1})^2(y+\frac{c'-2}{gn_1})},$$

where

$$\begin{aligned}
&Q(n_1, x, y, T') \\
&\quad = \frac{1}{g}\left\{\left(x + \frac{a'+1}{n_1}\right)\left(1 - y + \frac{d'+1}{gn_1}\right)\right. \\
&\qquad \times \left(1 - x + gy + \frac{b'+c'-3}{n_1}\right) \\
&\qquad + \left(1 - x + \frac{b'-2}{n_1}\right)\left(y + \frac{c'-2}{gn_1}\right) \\
&\qquad \left. \times \left(x + g(1 - y) + \frac{a'+d'-1}{n_1}\right)\right\}.
\end{aligned}$$

We have

$$\frac{n_1^{k/2}\mathbf{E}_T[l_{\psi_e}^k(T'')]}{2^k z_{1-\alpha/2}^k} = \sum_{x,y\in\mathcal{F}_{n_1}} \{n_1 \operatorname{var}(\psi_e|T'')\}^{k/2} p_{Z_{n_1,T'}}(x,y|n_1,n_0,T')$$

$$= \frac{\mathbf{B}(a',b'-k_{\psi_e})\mathbf{B}(c'-k_{\psi_e},d')}{\mathbf{B}(a',b')\mathbf{B}(c',d')} \times \sum_{x,y\in\mathcal{F}_{n_1}} h_{n_1}(x,y) p_{Z_{n_1,T^\star}}(x,y|n_1,n_0,T^\star)$$

$$= \frac{\mathbf{B}(a',b'-k_{\psi_e})\mathbf{B}(c'-k_{\psi_e},d')}{\mathbf{B}(a',b')\mathbf{B}(c',d')} \times \mathbf{E}_{Z_{n_1,T^\star}}\left[h_{n_1}\left(X_{n_1,T^\star},Y_{n_1,T^\star}\right)\right],$$

where $T^\star = (a', b'-k_{\psi_e}, c'-k_{\psi_e}, d')$ and

$$h_{n_1}(x,y) = \frac{\mathbf{B}(a',b')\mathbf{B}(c',d')}{\mathbf{B}(a',b'-k_{\psi_e})\mathbf{B}(c'-k_{\psi_e},d')} \times \{n_1 \operatorname{var}(\psi_e|T'')\}^{k/2} \frac{p_{Z_{n_1,T'}}(x,y|n_1,n_0,T')}{p_{Z_{n_1,T^\star}}(x,y|n_1,n_0,T^\star)}.$$

The proof is completed if we can show that

$$\lim_{n_1\to\infty} \mathbf{E}_{Z_{n_1,T^\star}}\left[h_{n_1}\left(X_{n_1,T^\star},Y_{n_1,T^\star}\right)\right] = \mathbf{E}_{Z^\star}[h(X^\star,Y^\star)],$$

where $Z_{n_1,T^\star} \to^d Z^\star = (X^\star, Y^\star)$ and $\lim_{n_1\to\infty} h_{n_1}(x,y) = h(x,y)$. We make use of theorem 25.12 of Billingsley (1995, p. 338):

1. Define $Z^\star = (X^\star, Y^\star)$, where $X^\star \sim \text{Be}(a', b'-k_{\psi_e})$ and $Y^\star \sim \text{Be}(c'-k_{\psi_e}, d')$ are independent. The first step (i.e., $Z_{n_1,T^\star} \to^d Z^\star$), is a straightforward application of Theorem B.1.
2. Next, define

   $$c_{n_1}\left(k,k_{\psi_e},T'\right) = \frac{g^{k_{\psi_e}} n_1^{2k_{\psi_e}} \Gamma(n_1+a'+b'-k_{\psi_e})\Gamma(gn_1+c'+d'-k_{\psi_e})}{\Gamma(n_1+a'+b')\Gamma(gn_1+c'+d')}.$$

   We have

   $$h_{n_1}(x,y) = c_{n_1}\left(k,k_{\psi_e},T'\right)[Q(n_1,x,y,T')]^{k/2} \times u_{n_1}\left(x,k,k_{\psi_e},a',b'\right)u_{gn_1}\left(1-y,k,k_{\psi_e},d',c'\right),$$

   where

   $$u_{n_1}(x,k,k_{\psi_e},e,f) = \left(\frac{e}{n_1}+x\right)^{k/2} \frac{\prod_{i=1}^{k_{\psi_e}}\left(\frac{f-i}{n_1}+1-x\right)}{\left(\frac{f-1}{n_1}+1-x\right)^k\left(\frac{f-2}{n_1}+1-x\right)^{k/2}}$$

   with $\lim_{n_1\to\infty} Q(n_1,x,y,T') = \frac{x(1-x)}{g} + y(1-y)$, $\lim_{n_1\to\infty} u_{n_1}(x,k,k_{\psi_e},a',b') = x^{k/2}(1-x)^{k_{\psi_e}-3k/2}$, and $\lim_{n_1\to\infty} c_{n_1}(k,k_{\psi_e},T') = 1$. Therefore, $\lim_{n_1\to\infty} h_{n_1}(x,y) = x^{k/2}(1-y)^{k/2}(1-x)^{k_{\psi_e}-3k/2}y^{k_{\psi_e}-3k/2}[\frac{x(1-x)}{g}+y(1-y)]^{k/2} = h(x,y)$.
3. This step involves showing that the sequence of functions $h_{n_1}(x,y)$ is uniformly bounded. Thus the sequence of random variables $h_{n_1}(X_{n_1,T^\star}, Y_{n_1,T^\star})$ is uniformly integrable according to an obvious extension of exercise 25.8 of Billingsley (1995, p. 340) to more than one argument. First, note that $\|u_{n_1}(x,k,k_{\psi_e},e,f)\| \le (\frac{e}{n_1}+1)^{k/2}\sqrt{\frac{f-k_{\psi_e}}{n_1}+1}$ and $\|c_{n_1}(k,k_{\psi_e},T')\| \le 1$. Second, it is also easily seen that the function $Q(n_1,x,y,T')$ is uniformly bounded. Therefore, the functions $h_{n_1}(x,y)$ are uniformly bounded.

Hence

$$\lim_{n_1\to\infty} \frac{n_1^{k/2}\mathbf{E}_T[l_{\psi_e}^k(T'')]}{2^k z_{1-\alpha/2}^k} = \frac{\mathbf{B}(a',b'-k_{\psi_e})\mathbf{B}(c'-k_{\psi_e},d')}{\mathbf{B}(a',b')\mathbf{B}(c',d')} \int_0^1\int_0^1 h(x,y)p(x,y)\,dx\,dy,$$

which completes the proof.

[Received September 2003. Revised July 2005.]

REFERENCES

Adcock, C. J. (1987), "A Bayesian Approach to Calculating Sample Sizes for Multinomial Sampling," *The Statistician*, 36, 155–159; Corr., 37, 239.

——— (1988), "A Bayesian Approach to Calculating Sample Sizes," *The Statistician*, 37, 433–439.

——— (1993), "An Improved Bayesian Procedure for Calculating Sample Sizes in Multinomial Sampling," *The Statistician*, 42, 91–95.

——— (1997), "Sample Size Determination: A Review," *The Statistician*, 46, 261–283.

Ashby, D., Hutton, J. L., and McGee, M. A. (1993), "Simple Bayesian Analyses for Case-Control Studies in Cancer Epidemiology," *The Statistician*, 42, 385–397.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Billingsley, P. (1995), *Probability and Measure* (3rd ed.), New York: Wiley.

Carlin, J. B. (1992), "Meta-Analysis for 2 × 2 Tables: A Bayesian Approach," *Statistics in Medicine*, 11, 141–158.

Chen, K. (2000), "Optimal Sequential Designs of Case-Control Studies," *The Annals of Statistics*, 28, 1452–1471.

Chen, M.-H., and Shao, Q.-M. (1999), "Monte Carlo Estimation of Bayesian Credible and HPD Intervals," *Journal of Computational and Graphical Statistics*, 8, 69–92.

Cornfield, J. (1951), "A Method of Estimating Comparative Rates From Clinical Data, With Applications to Cancer of the Lung, Breast and Cervix," *Journal of the National Cancer Institute*, 11, 1269–1275.

De Santis, F., and Perone Pacifico, M. (2001), "Sample Size Determination for Interval Estimation," Technical Report 4, Università di Roma, La Sapienza, Dipartimento di Statistica, Probabilità Statistiche Applicate,

De Santis, F., Perone, P. M., and Sambucini, V. (2004), "Optimal Predictive Sample Size for Case-Control Studies," *Applied Statistics*, 53, 427–441.

Dharmadhikari, S., and Joag-Dev, K. (1988), *Unimodality, Convexity, and Applications*, Boston: Academic Press.

Gail, M., Williams, R., Byar, D. P., and Brown, C. (1976), "How Many Controls?" *Journal of Chronic Diseases*, 29, 723–732.

Gardner, M. J., and Altman, D. G. (1986), "Confidence Intervals Rather Than *p* Values: Estimation Rather Than Hypothesis Testing," *British Medical Journal*, 292, 746–750.

Ghosh, M., and Chen, M.-H. (2002), "Bayesian Inference for Matched Case-Control Studies," *Sankhyā*, 64, 107–127.

Hashemi, L., Nandram, B., and Goldberg, R. (1997), "Bayesian Analysis for a Single 2 × 2 Table," *Statistics in Medicine*, 16, 1311–1328.

Hyndman, R. J. (1996), "Computing and Graphing the Highest-Density Region," *The American Statistician*, 50, 120–126.

Joseph, L., and Bélisle, P. (1997), "Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means," *The Statistician*, 46, 209–226.

Joseph, L., du Berger, R., and Bélisle, P. (1997), "Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination," *Statistics in Medicine*, 16, 769–781.

Joseph, L., and Wolfson, D. B. (1997), "Interval-Based versus Decision-Theoretic Criteria for the Choice of Sample Size," *The Statistician*, 46, 145–149.

Joseph, L., Wolfson, D. B., and du Berger, R. (1995), "Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals," *The Statistician*, 44, 143–154.

Latorre, G. (1982), "The Exact Posterior Distribution of the Cross-Ratio of a 2 × 2 Contingency Table," *Journal of Statistical Computation and Simulation*, 16, 19–24.

——— (1984), "Bayesian Inference in 2 × 2 and 2 × 2 × 2 Contingency Tables," *Metron*, 42, 169–184.

Lemeshow, S., Hosmer, D. W., and Stewart, J. P. (1981), "A Comparison of Sample Size Determination Methods in the Two-Group Trial Where the Underlying Disease Is Rare," *Communications in Statistics, Part B— Simulation and Computation*, 10, 437–449.

Lindley, D. V. (1997), "The Choice of Sample Size," *The Statistician*, 46, 129–138.

M'Lan, C. E. (2002), "Bayesian Sample Size Calculations for Cohort and Case-Control Studies," unpublished doctoral thesis, McGill University.

Marrie, R. A., Wolfson, C., Sturkenboom, M. C., Gout, O., Heinzlef, O., Roullet, E., and Abenhaim, L. (2000), "Multiple Sclerosis and Antecedent Infections: A Case-Control Study," *Neurology*, 54, 2307–2310.

Marshall, R. J. (1988), "Bayesian Analysis of Case-Control Studies," *Statistics in Medicine*, 7, 1223–1230.

Müller, P. (1999), "Simulation-Based Optimal Design," in *Bayesian Statistics 6*, eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 459–474.

Müller, P., and Parmigiani, G. (1995), "Optimal Design via Curve Fitting of Monte Carlo Experiments," *Journal of the American Statistical Association*, 90, 1322–1330.

Müller, P., and Roeder, K. (1997), "A Bayesian Semiparametric Model for Case-Control Studies With Errors in Variables," *Biometrika*, 84, 523–537.

Nam, J.-M., and Fears, T. R. (1992), "Optimum Sample Size Determination in Stratified Case-Control Studies With Cost Considerations," *Statistics in Medicine*, 11, 547–556.

Nurminen, M., and Mutanen, P. (1987), "Exact Bayesian Analysis of Two Proportions," *Scandinavian Journal of Statistics*, 14, 67–77.

O'Neill, R. T. (1984), "Sample Sizes for Estimation of the Odds Ratio in Unmatched Case-Control Studies," *American Journal of Epidemiology*, 120, 145–153.

Pham-Gia, T., and Turkkan, N. (1992), "Sample Size Determination in Bayesian Analysis," *The Statistician*, 41, 389–397.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994), "Non-Hierarchical Logistic Models and Case-Only Designs for Assessing Susceptibility in Population-Based Case-Control Studies," *Statistics in Medicine*, 13, 153–162.

Satten, G. A., and Kupper, L. L. (1990), "Sample Size Requirements for Interval Estimation of the Odds Ratio," *American Journal of Epidemiology*, 131, 177–184.

Schlesselman, J. J. (1982), *Case-Control Studies: Design, Conduct, Analysis*, New York: Oxford University Press.

Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.), New York: Springer-Verlag.

Wang, F., and Gelfand, A. E. (2002), "A Simulation-Based Approach to Bayesian Sample Size Determination for Performance Under a Given Model and for Separating Models," *Statistical Science*, 17, 193–208.

Wickramaratne, P. J. (1995), "Sample Size Determination in Epidemiologic Studies," *Statistical Methods in Medical Research*, 4, 311–337.

Zelen, M., and Parker, R. A. (1986), "Case-Control Studies and Bayesian Inference," *Statistics in Medicine*, 5, 261–269.