Statistica Sinica **16**(2006), 697-719

DATA-ADAPTIVE SEQUENTIAL DESIGN FOR CASE-CONTROL STUDIES

Malay Ghosh and Bhramar Mukherjee

University of Florida

Abstract: Case-control study designs are popular in epidemiological research for their cost saving and time saving properties. The efficiency of the design depends on the choice of case-control ratio, which is often arbitrarily chosen, resulting in a loss of efficacy. We study sequential case-control designs where cases occur sequentially over time and propose a sampling rule and a simple Bayes stopping rule which lead to the optimal sequential case-control design. This sampling rule can be applied to "case-control within cohort" studies where controls are sampled from failure-free members of the cohort at each distinct failure time when a case occurs, the study design itself being intrinsically sequential in nature. The proposed stopping rule is shown to be first order asymptotically optimal. Simulation results indicate that finite sample performance of the stopping rule and the estimation rule is satisfactory for moderate sample sizes.

Key words and phrases: Asymptotically pointwise optimal (APO) rules, case-control ratio, cost effective, Laplace approximation, Nurse's health study, Synthetic case-control studies.

1. Introduction

Case-control designs have become increasingly popular in studying etiology of a rare disease since first conceived in the 1920's. Classical fixed sample case-control studies select separate samples from the case and control populations with the two sample sizes being fixed, and often the choice of the sample sizes are ad hoc. Fixed case-control studies usually cannot attain full efficiency as the optimal case-control ratio is unknown prior to a study. Moreover, for rare diseases, when the cases are occurring sequentially over time, fixed sample designs are usually inferior to sequential procedures in terms of time saving considerations. In practice, information on cases and controls are indeed collected sequentially in many case-control studies. See, for example, Vessey and Doll (1968), Boston Collaborative Drug Surveillance Project (1973), O'Neil and Anello (1978), Pasternack and Shore (1981) and O'Neill (1998). In certain investigations, there are ethical reasons for an early stopping of the study, such as saving samples or needing immediate public health policy actions (O'Neill (1998)). A sequential strategy is definitely a superior choice to a fixed sample size design when interim analysis and early stopping are possible options. The need for data adaptive sequential designs is even greater for individually matched case-control study with cases occurring sequentially over time where one must decide on the appropriate choice of controls to match a newly available case. In such study designs, a sequential sampling and stopping strategy could lead to substantial saving of resources.

For fixed sample size case-control studies, there is substantial literature on determining the sample size, which is an important aspect in designing any comparative experiment. Schlesselman (1974) provided a sample size formula based on Cochran's (1954) statistic. Munoz and Rosner (1994) study sample size determination for the Mantel-Haenszel test (1959) on stratified data when both sets of marginals considered in each table are fixed. Woolson, Bean and Rojas (1986) propose sample size formulas for Cochran's statistic in stratified and unstratified case-control analyses. Nam (1992) derives a sample size formula for Cochran's statistic with continuity correction, and Nam and Fears (1990, 1992a, b) determine sample size for stratified case-control studies with cost per control varying by strata. All these papers consider a dichotomous covariate of the nature: exposure and non-exposure. Nam (1997) and Lui (1990, 1993) consider an extension to multiple risk factors.

O'Neill and Anello (1978) were the first ones to indicate the advantage of sequential design in confirmatory trials. Several published studies on the association between breast cancer and a widely used anti-hypertensive drug reserpine were deemed to be inconclusive. Motivated by this example, O'Neill and Anello (1978) proposed a simple stopping rule based on Wald's SPRT for a matched case-control study with a single binary covariate and fixed matching ratio. The problem of how many controls should be matched with a case was not addressed in their framework.

Chen (2000) provides an interesting example of where a sequential strategy may be the best choice. Consider the Nurses Health Study Stampfer, Willet, Coldits, Roser, Speizer and Henekens (1985). At the onset of the study, blood serum samples for 100,000 subjects were frozen for subsequent use. In later phases of the study, some subjects developed coronary disease and others remained disease free. Investigators wanting to study the effect of serum vitamins A and F on the risk of disease were only authorized to use a maximum of 2,000 frozen samples due to high cost of laboratory analysis and small amount of stored serum for each individual. The question then is how one should allocate case or control samples to such time as 2,000 samples are taken, so that the final estimate of the relative risk parameters would be most accurate. Could a specified accuracy of estimation be achieved without using 2,000 serum samples? Chen (2000) proposes a sequential sampling strategy based on extending asymptotic normality of the semiparametric MLE from fixed sample size to sequential designs. A stopping rule is proposed when one wants a fixed width confidence interval for the relative risk parameter.

Sequential principles may naturally be called for in a "synthetic" case-control design (Whittemore (1981) and Prentice (1986)), as a means of reducing the number of subjects for whom covariate data need to be assembled in the context of a cohort study. For (instantaneous) relative risk estimation, this method requires matching each subject experiencing a failure (a case) with a desirable number of subjects who are failure-free (controls) at the failure time when the case occurs. Since the cases naturally occur over time in this framework, in order to determine the "desirable" number of controls at each step, a sequential optimizing principle may be adopted as discussed in the current paper.

We look at the sequential design problem for case control studies from a Bayesian perspective, furnishing a sampling rule as well as a simple, easy-to-use stopping rule. The sampling rule we propose is akin to the one proposed by Chen (2000). However in this paper, our main focus is an easy-to-use stopping rule, accompanying a sampling rule. To this end, we consider sequential Bayes stopping rules. Sequential Bayes analysis is primarily concerned with two problems: (1) when to stop taking observations, and (2) what to use as a decision when stopped. It is well-known that in a Bayesian framework, regardless of the stopping time, the optimal decision rule given the stopped sequence is the Bayes rule with respect to the posterior distribution at that time. Thus in a Bayesian framework, the main problem is to find the optimal Bayes stopping time. The computations involved in obtaining a Bayesian stopping rule by backward induction could be formidable. Consequently, finding good approximations to optimal Bayes stopping rules is of prime importance and many candidates have been proposed. In particular, Bickel and Yahav (1967, 1968, 1969) proposed a class of rules termed *Asymptotically Pointwise Optimal* (A.P.O.) rules. Woodroofe (1981) and Rehalia (1984) established that these rules are non-deficient in the sense that asymptotically they have the same Bayes risk as the optimal Bayes procedure.

Sequential principles have often been exploited to develop easy-to-use strategies for applied scientists. APO rules have also been used effectively in solving problems like the detection of influenza epidemics (Baron (2002)). Dalal and Mallows (1988) propose APO rules for stopping the testing of software prior to release, while considering the trade-off between cost of continued testing and the expected losses due to any bugs remaining in the released code. Fakhre-Zakeri and Slud (1996) consider APO rules for sequential size-dependent searches and apply it to software reliability testing. The potential of applications of APO rules has not yet been fully explored. As we will note, the APO rules perform quite well in the case-control context though the analysis is non-conjugate and remains outside the one parameter exponential family framework, the cases for which most of the theoretical optimality properties have been developed.

The rest of the paper is organized as follows. In Section 2, we describe the prospective logistic regression model for analyzing case-control data. We specify prior distributions on the regressor related parameters and describe the approximate asymptotic distribution of the posterior estimate for these parameters. In Section 3, we introduce the sampling rule for choosing cases and controls at each stage as well as the stopping rule, and prove the first order optimality of the proposed stopping rule. In Section 4, we illustrate the sequential procedure in detail with a single binary exposure. We conduct a small scale simulation study for the binary exposure case to judge the finite sample performance of the proposed estimation strategy. Section 5 contains a brief outline toward possible extension of this method to a group-sequential framework, and for matched case-control studies.

2. Likelihood, Prior and Posterior

For simplicity, we consider a disease variable Y and a single exposure X. The discussion and the results generalize in a straightforward way to a vector of multiple exposures. The data at stage n is denoted by $\mathcal{D}_n = \{(x_i, y_i)\colon\ i = 1, \ldots, n\}$. Let n_1 denote the number of cases at stage n and take $r_n = n_1/n$. Let $h(t) = \exp(t)/(1+\exp(t))$. Then a prospective logistic regression model for disease incidence is

$$P(Y = 1|X = x) = h(\gamma + \beta x). \tag{1}$$

We denote by $\phi_1(x)$ and $\phi_0(x)$ the densities of X in the case and control populations, respectively. We assume that $\phi_1(x)$ and $\phi_0(x)$ comply with Prentice-Pyke type conditions (Prentice and Pyke (1979)) to ensure the validity of the model. Then the usual prospective logistic likelihood, conditional on the x_i values is:

$$L_n(\gamma, \beta) = \prod_{i=1}^{n} [h(\gamma + \beta x_i)]^{y_i} [1 - h(\gamma + \beta x_i)]^{1-y_i}. \tag{2}$$

Hence the log-likelihood $l_n(\gamma, \beta)$ and the observed Fisher information matrix $\mathbf{I}_n(\gamma, \beta)$ are given by

$$l_n(\gamma, \beta) = \sum_{i=1}^{n} [y_i(\gamma + \beta x_i) - \log(1 - h(\gamma + \beta x_i))], \tag{3}$$

$$\mathbf{I}_n(\gamma, \beta) = \begin{bmatrix} \sum_{i=1}^n h'(\gamma + \beta x_i) & \sum_{i=1}^n x_i h'(\gamma + \beta x_i) \\ \sum_{i=1}^n x_i h'(\gamma + \beta x_i) & \sum_{i=1}^n x_i^2 h'(\gamma + \beta x_i) \end{bmatrix}, \tag{4}$$

where $h'(t) = \exp(t)/(1+\exp(t))^2$ is the derivative of $h(t)$. Note also that in view of (4), $\partial^{j+k} l_n(\gamma,\beta)/\partial\gamma^j\partial\beta^k$ $(j,k = 0,1,\ldots; j+k \geq 3)$, do not involve the y_i.

Let $\boldsymbol{\eta} = (\gamma,\beta)^T$ denote the true parameter vector, and let $\widehat{\boldsymbol{\eta}}_n = (\widehat{\gamma}_n, \widehat{\beta}_n)^T$ denote the vector of maximum likelihood estimates of $\boldsymbol{\eta}$ at stage n. That is, $(\widehat{\gamma}_n, \widehat{\beta}_n)$ is the unique solution to the system of equations

$$\sum_{i=1}^{n} \Big[\binom{1}{x_i} (y_i - h(\gamma + \beta x_i)) \Big] = 0. \tag{5}$$

Then from the asymptotic normality of the MLE, it follows that, for given values of x,

$$\mathbf{I}_n^{1/2}(\boldsymbol{\eta})(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{d} N_2(0, \mathbf{I}). \tag{6}$$

Then from (4) and (6), and plugging in the estimates for γ and β, it follows that, for given values of x_i and a given n, an estimate of $\mathrm{Var}\,(\widehat{\beta_n})$ is

$$\widehat{\mathrm{Var}\,(\widehat{\beta_n})} \approx \frac{\sum_{i=1}^{n} h'(\widehat{\gamma}_n + \widehat{\beta}_n x_i)}{|\mathbf{I}_n(\hat{\boldsymbol{\eta}}_n)|}. \tag{7}$$

Recall that we have assumed $P(X|Y = d) = \phi_d(x)$, $d = 0,1$. Chen (2000) extended the asymptotics for the semiparametric MLE of β for a given n to the sequential context, and established that

$$n\widehat{\mathrm{Var}\,(\widehat{\beta_n})} \longrightarrow [\Sigma(r)]^{-1} \qquad \text{as } r_n \to r \text{ in probability}, \tag{8}$$

where

$$\Sigma(r) = r\Sigma_1(r) + (1-r)\Sigma_0(r), \tag{9}$$

$$\Sigma_1(r) = E_1[(X - A(r))^2 h'(\gamma^*(r) + \beta X)], \tag{10}$$

$$\Sigma_0(r) = E_0[(X - A(r))^2 h'(\gamma^*(r) + \beta X)], \tag{11}$$

$$A(r) = \frac{rE_1(Xh'(\gamma^*(r) + \beta X)) + (1-r)E_0(Xh'(\gamma^*(r) + \beta X))}{[rE_1(h'(\gamma^*(r) + \beta X)) + (1-r)E_0(h'(\gamma^*(r) + \beta X))]}. \tag{12}$$

Here E_d denotes expectation on X under the density $\phi_d, d = 0,1$, and

$$\gamma^*(r) = \log(\frac{r}{1-r}) - \log(\frac{P(Y=1)}{P(Y=0)}) + \gamma. \tag{13}$$

Proposition 2.1 of Chen (2000) establishes the above result, using the relationship between prospective logistic model and retrospective likelihood at several steps of the proof.

We adopt a Bayesian approach and assume a bivariate normal prior on η, $\pi(\boldsymbol{\eta}) \sim N_2(\boldsymbol{m}, \boldsymbol{W})$. Then the posterior for $\boldsymbol{\eta}$ is given by

$$\pi(\boldsymbol{\eta}|\mathcal{D}_n) = \frac{\exp[l_n(\boldsymbol{\eta})]\pi(\boldsymbol{\eta})}{\int \exp[l_n(\boldsymbol{\eta})]\pi(\boldsymbol{\eta})d\boldsymbol{\eta}}. \tag{14}$$

For convenience of writing, we now let $\eta_1 = \gamma$, $\eta_2 = \beta$, $\hat{\eta}_{1n} = \hat{\gamma}_n$ and $\hat{\eta}_{2n} = \hat{\beta}_n$. By expanding $l_n(\boldsymbol{\eta})$ around the MLE $\hat{\boldsymbol{\eta}}_n$, and noting that $\nabla l_n(\hat{\boldsymbol{\eta}}_n) = \mathbf{0}$, we have

$$\begin{aligned} &\exp[l_n(\boldsymbol{\eta})] \\ &= \exp\Big[l_n(\hat{\boldsymbol{\eta}}_n) - \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)^T \boldsymbol{I}_n(\hat{\boldsymbol{\eta}}_n)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n) + K_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n) + R_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)\Big] \\ &= \exp[l_n(\hat{\boldsymbol{\eta}}_n)] \exp[-\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)^T \boldsymbol{I}_n(\hat{\boldsymbol{\eta}}_n)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)]\Big(1 + K_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n) + R_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)\Big), \end{aligned} \tag{15}$$

where

$$K_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n) = \frac{1}{6} \sum_{k,l,m=1,2} (\eta_k - \hat{\eta}_{nk})(\eta_l - \hat{\eta}_{nl})(\eta_m - \hat{\eta}_{nm}) \frac{\partial^3 l_n(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_l \partial \eta_m}\Big|_{\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}_n}, \tag{16}$$

and $R_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)$ denotes the remainder terms involving fourth and higher order partial derivatives. By (14) and (15), we have

$$\pi(\boldsymbol{\eta}|\mathcal{D}_n) = \frac{B_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)}{\int B_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)d\boldsymbol{\eta}}, \tag{17}$$

where

$$B_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n) = \exp[-\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)^T \boldsymbol{I}_n(\hat{\eta}_n)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)]\pi(\boldsymbol{\eta})\Big(1 + K_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n) + R_n(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}_n)\Big). \tag{18}$$

Based on (17) and (18), we now have the following lemma.

Lemma 1. *Assume that* $\lim_{n\to\infty} \inf n^{-2} \sum_i \sum_j (x_i - x_j)^2 h'(\gamma + \beta x_i) h'(\gamma + \beta x_j) > 0$, *a.s., and* $\sum_{i=1}^n x_i^2 = O_p(n)$. *Then*
(i) $\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\eta}}_n) = O_p(n^{-1})$;
(ii) $E(\boldsymbol{\eta}|\mathcal{D}_n) = \hat{\boldsymbol{\eta}}_n + O_p(n^{-1})$;
(iii) $E[(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)^\top|\mathcal{D}_n] = \boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\eta}}_n) + O_p(n^{-3/2})$.

Proof. The proof of Lemma 1 is technical, and deferred to the supplementary materials available online.

It may be noted that as a consequence of (ii) and (iii) of Lemma 1,

$$\mathrm{Var}\,(\boldsymbol{\eta}|\mathcal{D}_n) = \mathrm{Var}\,(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n|\mathcal{D}_n) = \boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\eta}}_n) + O_p(n^{-\frac{3}{2}}). \tag{19}$$

The following result, which is immediate from (19), will be used in the subsequent sections:

$$n\,\mathrm{Var}\,(\boldsymbol{\beta}|\mathcal{D}_n) = n\,\widehat{\mathrm{Var}\,(\hat{\beta}_n)} + O_p(n^{-\frac{1}{2}}), \tag{20}$$

where $\widehat{\mathrm{Var}\,(\hat{\beta}_n)}$ was defined at (7).

3. The Sequential Design

In this section, we derive the sampling strategy at each stage and propose an accompanying easy-to-use stopping rule. We establish certain large sample properties of our design. Before we begin describing the proposed design, we point out that the design question addressed in this paper is very different from the classical problem of designing covariates x_i in a logistic regression model, for which there exists a number of papers (Chaloner and Larntz (1989) and Wu (1985)). Our goal is to select the "responses" or the cases and controls in a data-adaptive, sequential way, to optimally estimate the odds ratio parameter $\exp(\beta)$. Designing the covariates is not a meaningful proposition in the context of a retrospective case-control study.

3.1. The Sampling Rule

We noted in Lemma 1 that the posterior variance of β for any n is asymptotically equivalent to the asymptotic variance of the ML estimate of β, namely $\mathrm{Var}\,(\hat{\beta}_n)$, n times which converges to $[\Sigma(r)]^{-1}$.

Following Chen (2000), it can be shown that $[\Sigma(r)]^{-1}$ is a *strictly convex* function of r on [0,1] with a unique minimum achieved at r^*, say. A multiple of the derivative of $[\Sigma(r)]^{-1}$ is given by

$$D(r) = \frac{r}{1-r}\Sigma_1(r) + \frac{1-r}{r}\Sigma_0(r). \tag{21}$$

The optimal choice for the case-control ratio r^* that minimizes $[\Sigma(r)]^{-1}$ is the unique solution to the equation $D(r) = 0$. An estimator of $D(r)$, say $\widehat{D}(r)$, can be obtained by plugging in current estimates of γ, β and the empirical analogues of ϕ_1 and ϕ_0 in (21). Chen (2000) also justifies the use of $\hat{\gamma}_n$ as a working estimate of $\hat{\gamma}^*(r_n)$ (as defined in (13) and also appearing in $\hat{D}(r_n)$) at each step.

Result (Chen (2000)): A sequential case-control design with case percentage r_n is asymptotically efficient if and only if $r_n \to r^*$ as $n \to \infty$.

The proposed sampling rule at the nth stage of the experiment is as follows:

- take a sample from the case population if $\widehat{D}(r_n)<0$;
- draw a sample from the control population if $\widehat{D}(r_n)>0$;
- choose arbitrarily if $\widehat{D}(r_n)=0$.

The intuitive idea behind the sampling rule is that, since we are looking for the minimum of a convex function of r_n, if the estimated derivative is positive we decrease r_n (i.e., choose a control sample) to move closer to the minimum and, conversely, when the derivative is negative we increase r_n (i.e., choose a case sample) to achieve the minimum.

Note that this sampling rule resembles the stochastic approximation ideas of Wu (1985), where the goal is to approximate the root of a system of appropriate equations sequentially. Here our goal is not to find the exact value of r^*, but to accurately estimate β.

Chen (2000) also established that under such a sampling scheme, the case-control sampling is asymptotically efficient in the sense that

$$r_n \to r^* \text{ a.s. as } n \to \infty. \tag{22}$$

He proposed a stopping rule based on a fixed width confidence interval for β. In the following, we propose a very simple stopping rule when estimation is done in a Bayesian framework.

3.2. The APO stopping rule

We begin with the general framework for A.P.O. stopping rules. For a detailed formulation the reader is referred to the papers by Bickel and Yahav (1967, 1968) and other references mentioned in the introduction.

Let the loss function be $L_n(c) = Y_n + nc$, with $P(Y_n > 0) = 1$ and $Y_n \to 0$ as $n \to \infty$. One may think of Y_n as the posterior risk under a prior at time n, with c as the cost per sampling unit. Let $\mathcal{T}$ be the class of all stopping rules. A stopping rule T is defined to be A.P.O. if, for any $S \in \mathcal{T}$,

$$\limsup_{c\to 0} \frac{L_T(c)}{L_S(c)} \le 1 \qquad \text{a.s.}.$$

In many standard situations the A.P.O. rule turns out to be of the form $N = \inf\{n(\ge m) : n \ge (Y_n/c)\}$, where m is the initial sample size.

In the case-control sampling situation where the parameter of interest is β, we take

$$L_n(c) = \text{Var}\,(\beta|\mathcal{D}_n) + cn, \tag{23}$$

where c is the cost per unit sample. Let $G_n = n\text{Var}\,(\beta|\mathcal{D}_n)$. Then we propose the following APO stopping time:

$$N = \inf\{n(\ge m) : n \ge (\frac{G_n}{c})^{\frac{1}{2}}\}, \tag{24}$$

where m is the initial sample size. Also, $N \to \infty$ a.s. as $c \to 0$.

Remark 1. Chen (2000) proposed a stopping time N_d associated with a fixed width confidence interval $(\hat{\beta}_N - d, \hat{\beta}_N + d)$ with confidence coefficient $1 - \alpha$ that is asymptotically consistent and efficient under the sampling rule as described in Section 3.1. The proposed stopping time is of the form

$$N_d = \inf\left\{n \geq 1 : n \geq \frac{z^2_{\frac{\alpha}{2}}}{d^2 \hat{\Sigma}(\hat{r}^*_n)}\right\},$$

where $\hat{r}^*_n$ is the solution to $\hat{D}(r) = 0$ at stage n (the construction of $\hat{D}(r)$ is described right after(21)), and z_α is the upper $100(1-\alpha)$th percentile of the standard normal distribution.

Our stopping rule cannot be directly compared with the fixed width confidence interval based stopping rule of Chen (2000), as our goal is to maximize the precision of the point estimate of β. The two objectives are quite different. The development of A.P.O rules related to set estimation can be found in Gleser and Kunte (1976). They consider a loss function that is a linear combination of the length of the interval, the indicator of non-coverage, and the cost of sampling. Again, this is which is not directly comparable to the fixed-width confidence interval based approach discussed in Chen (2000).

Another important aspect of sequential inference is sequential testing of hypotheses with a desired level of significance, and the power to detect specified effect sizes. This leads to the more common framework of deriving power curves for given effect sizes. Chen (2000) justifies the use of the sampling rule as discussed in the current paper in conjunction with the classical theory of sequential tests (Siegmund (1985)) for case-control sampling. Possible uses of A. P. O. rules in the context of testing of hypotheses are indicated in Bickel and Yahav (1967).

Remark 2. In the example cited in the introduction, the Nurse's Health Study (1985), there was an upper bound of 2,000 on the number of allowable samples. In such instances, the stopping rule would be modified to $\min(N, 2000)$. Similar modification will be needed if there are only a finite number of cases available, which is often the case for a rare disease.

Remark 3. One may naturally question the structure of the loss function $L_n(c)$, where c and $\mathrm{Var}\,(\beta|\mathcal{D}_n)$ may vary on quite different scales. However, in our set-up, one can calibrate c as needed, to reach compatibility with the desired level of accuracy for estimating β. The large sample properties are in fact obtained as $c \to 0$ (implying $N \to \infty$ a.s.). To elicit the ideal utility surface empirically in a specific problem is an interesting issue in itself, and will depend critically on the application.

Remark 4. The framework can be extended to choose the optimal case-control ratio when one has a vector of multivariate exposures instead of a single exposure.

In fact, Chen (2000) extends the sampling scheme to the case where one is dealing with a general p-dimensional vector β. In this case $\Sigma(r)$, as defined in (9), turns out to be a matrix function of r, and there may not exist an r^* such that $\Sigma(r^*) \leq \Sigma(r)$ for all $r \in [0,1]$ in the sense of positive definiteness. To determine the "optimal" case control ratio, it then becomes necessary to choose a general criterion in terms of $E[(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top]$, and to minimize this as a function of r_n. One could define a fairly general criterion like

$$\text{trace}(\boldsymbol{B}^\top E[(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top]\boldsymbol{B}) \tag{25}$$

for a fixed $p \times p$ matrix $\boldsymbol{B} \neq \mathbf{0}$, for example. A typical choice of $\boldsymbol{B}$ is then $Diag(b_1, \ldots, b_p)$, where b_i can be viewed as a weight attached to the accuracy of estimating the effect of the ith component of the covariate vector. For $\boldsymbol{B} = \boldsymbol{I}$, the criterion reduces to the squared norm $E||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||^2$, so that interest lies in estimating the effect of the entire p-variate exposure vector.

In a typical case-control study, one is often interested in controlling or adjusting for the effect of other covariates like age, ethnicity, family history of the disease, and the like, while studying the association of the disease and a set of risk factors. In such situations, one could simply choose $\boldsymbol{B}$ as a diagonal matrix with the diagonal elements corresponding to the covariates set at zero, and the diagonal elements corresponding to the exposure variables of interest set at nonzero values. The multivariate framework is very general and can handle many situations and objectives by suitable selection of the form of the fixed matrix $\boldsymbol{B}$.

The intriguing observation in Chen (2000) is that the limiting value of the criterion function in (25) as $r_n \to r$, a.s., is $\text{trace}(\boldsymbol{B}^\top \Sigma^{-1}(r)\boldsymbol{B})$. Moreover, the convexity of this limiting function as a function of r is retained even in this fully multivariate set-up. The derivative function $D_{\boldsymbol{B}}(r)$ of the limiting expression can also be estimated at each step and, consequently, the sampling rule as described in the case of scalar β can be directly adapted to the multivariate setting.

As far as the A.P.O. stopping rule in the multiparameter setting is concerned, the loss function $L_n(c)$ can also be directly modified in accordance with a particular choice of the criterion function as described in (25). In the expression for $L_n(c)$, as given in (23), $\text{Var}\,(\beta|\mathcal{D}_n)$ can be replaced by the posterior expectation of the chosen criterion function given $\mathcal{D}_n$, and the MLE $\hat{\boldsymbol{\beta}}_n$ in (25) related by the Bayes estimate $E(\boldsymbol{\beta}|\mathcal{D}_n)$. Generalization of the A.P.O. rule to this setting has been described in Hoekstra (1989), and the structure of the A.P.O. rule is exactly similar to the case with a scalar β.

Next, returning to our set-up with a scalar β, we present the key intuitive idea behind the asymptotic behavior of this stopping rule, before presenting the formal theorem. Let

$$U_n = n\,\widehat{\text{Var}\,(\hat{\beta}_n)}. \tag{26}$$

- By Lemma 1 and (20), $G_n = U_n + O_p(n^{-1/2})$. By (8), the leading term $U_n \xrightarrow{P} [\Sigma(r)]^{-1}$ as $n \to \infty$ and, in fact, to $[\Sigma(r^*)]^{-1}$ as $r_N \to r^*$ under the given sampling rule (see (21)).
- Since $N \to \infty$ a.s. as $c \to 0$, $U_N \xrightarrow{P} [\Sigma(r^*)]^{-1}$ as $c \to 0$. Hence, $G_N \xrightarrow{P} [\Sigma(r^*)]^{-1}$ as $c \to 0$.

This argument leads to the fact that the combination of stopping rule and sampling rule lead to an efficient sampling scheme for estimating β. We now present the more formal theorem regarding the first order asymptotic optimality of the stopping rule.

Theorem 1. *For the stopping time N as defined in* (24),
(i) $P(N < \infty) = 1$;
(ii) $cN^2 \xrightarrow{P} [\Sigma(r^*)]^{-1}$ *as* $c \to 0$*;*
(iii) $L_N(c)/\rho(c) \xrightarrow{P} 1$ *as* $c \to 0$*, where* $\rho(c) = \inf_{S \in \mathcal{T}} E(L_S(c)) = 2c^{1/2}[\Sigma(r^*)]^{-1/2}$*;*
(iv) $E[L_N(c)]/\rho(c) \to 1$ *as* $c \to 0$*. The A.P.O. rule is first order efficient or asymptotically optimal* (*A.O.*)*.*

Proof. The proof of this theorem is rather technical and is relegated to the supplementary materials available online.

4. Example: Binary Exposure

Since the above discussion is quite general, we illustrate our methods explicitly for the common situation in which one has a binary exposure (like smoking habit, use of oral contraceptive, or family history of cancer). For the ith sample, let $X_i = 1(0)$ if a person is exposed (unexposed); $Y_i = 1(0)$ if a person has the disease or is a case (does not have the disease or is a control). The joint probability function of Y_i and X_i is (Cox (1972) and Zelen and Parker (1986))

$$f(y_i, x_i) = \frac{\exp(\lambda x_i + \gamma y_i + \beta y_i x_i)}{1 + \exp(\lambda) + \exp(\gamma) + \exp(\lambda + \gamma + \beta)}.$$

This leads to

$$f(x_i|y_i) = \frac{\exp[(\lambda + \beta y_i)x_i]}{1 + \exp[(\lambda + \beta y_i)]};$$

$$f(y_i|x_i) = \frac{\exp[(\gamma + \beta x_i)y_i]}{1 + \exp[(\gamma + \beta x_i)]}.$$

The parameter β can be expressed as

$$\begin{aligned}\exp(\beta) &= \frac{f(x_i=1|y_i=1)/f(x_i=0|y_i=1)}{f(x_i=1|y_i=0)/f(x_i=0|y_i=0)}\\ &= \frac{f(y_i=1|x_i=1)/f(y_i=0|x_i=1)}{f(y_i=1|x_i=0)/f(y_i=0|x_i=0)}\\ &= \frac{f(1,1)f(0,0)}{f(0,1)f(1,0)}.\end{aligned}$$

We are interested primarily in inference for β. At the nth stage let there be n_1 cases and n_0 controls. Let n_{11} denote the number of exposed cases, and n_{10} denote the number of exposed controls. So $r_n = n_1/n = 1 - n_0/n$; Let, $u(t) = h'(t) = \exp(t)/[1+\exp(t)]^2$ and $\overline{h}(t) = 1 - h(t)$. Thus $u(t) = h(t)\overline{h}(t)$. Our goal is to derive an expression for $\Sigma(r)$ for this particular example.

Remark 5. Note that in this context, $\phi_d(x)$ is Bernoulli with success probabilities $h(\lambda + d\beta)$, $d = 0, 1$. In this bivariate binary case, instead of going through the prospective formulation and then taking expectation with respect to the covariate densities $\phi_d(x)$, as shown in (8)$-$(13), one can work directly with the retrospective likelihood to derive an expression for the asymptotic variance of the MLE of β, namely, $\hat{\beta}_n$, and obtain identical results. In the following we show the equivalence of these two approaches.

Prospective Formulation. It is shown in the supplementary materials available online that, following equations (8)$-$(11),

$$\Sigma(r) = (1-r)\frac{h(\gamma^*(r)+\beta)h(\lambda)h(\gamma^*(r))\overline{h}(\lambda)}{h(\gamma^*(r)+\beta)h(\lambda)+h(\gamma^*(r))\overline{h}(\lambda)}. \tag{27}$$

Recall that as stated in (13), $\gamma^*(r) = \log(r/1-r) - \log(p_1/(1-p_1)) + \gamma$. For this problem, the marginal disease probability, $p_1 = P(Y=1)$, is easily calculated as

$$p_1 = P(Y=1) = \frac{\exp(\gamma)[1+\exp(\lambda+\beta)]}{1+\exp(\lambda)+\exp(\gamma)+\exp(\lambda+\gamma+\beta)}. \tag{28}$$

Hence,

$$\frac{p_1}{1-p_1} = \frac{\overline{h}(\lambda+\beta)}{\exp(\gamma)\overline{h}(\lambda)}. \tag{29}$$

Using the expression for $\gamma^*(r)$ we have,

$$h(\gamma^*(r)+\beta)h(\lambda) = \frac{\frac{r}{1-r}\frac{1-p_1}{p_1}\exp(\gamma+\beta)h(\lambda)}{1+\frac{r}{1-r}\frac{1-p_1}{p_1}\exp(\gamma+\beta)}. \tag{30}$$

From (29) and (30) we may write,

$$h(\gamma^*(r)+\beta)h(\lambda) = \frac{rh(\lambda+\beta)h(\lambda)}{rh(\lambda+\beta)+(1-r)h(\lambda)}, \tag{31}$$

$$h(\gamma^*(r))\overline{h}(\lambda) = \frac{r\overline{h}(\lambda+\beta)\overline{h}(\lambda)}{r\overline{h}(\lambda+\beta)+(1-r)\overline{h}(\lambda)}. \tag{32}$$

Plugging (31) and (32) in (27), it can be shown, after simplification, that

$$\begin{aligned}\Sigma^{-1}(r) &= \frac{1}{1-r}[h(\gamma^*(r)+\beta)h(\lambda)]^{-1} + [h(\gamma^*(r))\overline{h}(\lambda)]^{-1} \\ &= [rh(\lambda+\beta)\overline{h}(\lambda+\beta)]^{-1} + [(1-r)h(\lambda)\overline{h}(\lambda)]^{-1} \\ &= [ru(\lambda+\beta)]^{-1} + [(1-r)u(\lambda)]^{-1}. \end{aligned} \tag{33}$$

Next we illustrate that this analysis is exactly equivalent to using the retrospective likelihood directly, and using the asymptotic distribution for the MLE of β as derived from using the score function corresponding to the retrospective likelihood.

Direct use of retrospective likelihood. For this example, one can directly work with the likelihood function based on $f(x_i|y_i)$'s at the nth stage,

$$L_n(\lambda,\beta) = \frac{\exp[n_{11}(\lambda+\beta)+n_{10}\lambda]}{[1+\exp(\lambda+\beta)]^{n_1}[1+\exp(\lambda)]^{n_0}}.$$

The observed Fisher information matrix at the nth stage is

$$\mathbf{I}_n(\lambda,\beta) = n\begin{bmatrix} r_n u(\lambda+\beta)+(1-r_n)u(\lambda) & r_n u(\lambda+\beta) \\ r_n u(\lambda+\beta) & r_n u(\lambda+\beta)\end{bmatrix}.$$

If $r_n \to r$, a.s. $(0<r<1)$ as $n\to\infty$, then $n^{-1}\mathbf{I}_n$ converges a.s. to

$$\begin{bmatrix} ru(\lambda+\beta)+(1-r)u(\lambda) & ru(\lambda+\beta) \\ ru(\lambda+\beta) & ru(\lambda+\beta)\end{bmatrix}.$$

As before, let, $\hat{\lambda}_n$ and $\hat{\beta}_n$ denote the MLE's of λ and β. Then from the asymptotic normality of the MLE and the above expression for the Fisher Information matrix, $n^{1/2}(\hat{\beta}_n-\beta) \xrightarrow{d} N[0,g(\lambda,\beta,r)]$, where $g(\lambda,\beta,r) = [ru(\lambda+\beta)]^{-1} + [(1-r)u(\lambda)]^{-1}$. Note that $g(\lambda,\beta,r)$ is identical to $[\Sigma(r)]^{-1}$ as furnished in (33). This clearly demonstrates the equivalence of the two approaches.

Considering g (or $\Sigma^{-1}(r)$) as a function of r, one can immediately calculate

$$\frac{\partial}{\partial r}g(\lambda,\beta,r) = [u(\lambda)]^{-1}\frac{1}{(1-r)^2} - [u(\lambda+\beta)]^{-1}\frac{1}{r^2}, \tag{34}$$

$$\frac{\partial^2}{\partial r^2}g(\lambda,\beta,r) = [u(\lambda)]^{-1}\frac{2}{(1-r)^3} + [u(\lambda+\beta)]^{-1}\frac{2}{r^3}. \tag{35}$$

By (35), $g(\gamma,\beta,r)$ is a convex function of r, $0 < r < 1$, with minimum attained at

$$r^* = \frac{[u(\lambda)]^{\frac{1}{2}}}{[u(\lambda)]^{\frac{1}{2}} + [u(\lambda+\beta)]^{\frac{1}{2}}}. \tag{36}$$

Sampling Rule: Let $g'(\hat{\lambda}_n,\hat{\beta}_n,r_n) = \frac{\partial}{\partial r}g(\lambda,\beta,r)\Big|_{\lambda=\hat{\lambda}_n,\beta=\hat{\beta}_n,r=r_n}$. The exact expression for g' is obtained by plugging in $\lambda = \hat{\lambda}_n, \beta = \hat{\beta}_n$, $r = r_n$ in (34). At the nth stage of the experiment, take a sample from

- the case population if $g'(\hat{\lambda}_n,\hat{\beta}_n,r_n) < 0$;
- the control population if $g'(\hat{\lambda}_n,\hat{\beta}_n,r_n) > 0$;
- arbitrary if $g'(\hat{\lambda}_n,\hat{\beta}_n,r_n) = 0$.

Here $g'(\hat{\lambda}_n,\hat{\beta}_n,r_n)$ is identical to the estimated derivative of $[\Sigma(r)]^{-1}$, namely $\widehat{D(r_n)}$ as described in Section 3.1.

We assume the $N(\boldsymbol{m},\boldsymbol{W})$ prior on (λ,β). The posterior variance of β can then be calculated, as in Lemma 1, as $n\mathrm{Var}\,(\beta|\mathcal{D}_n) = G_n = U_n + O_p(n^{-1/2})$, where $U_n = n\widehat{\mathrm{Var}\,(\hat{\beta}_n)} = [(1-r_n)u(\hat{\lambda}_n)]^{-1} + [r_n\, u(\hat{\lambda}_n+\hat{\beta}_n)]^{-1}$.

Remark 6. In the light of Lemma 1, the APO rule N, as defined in (24), can again be approximated as (by replacing G_n with U_n in the definition of N)

$$\tilde{N} = \inf\{n \geq m : n \geq (\frac{U_n}{c})^{\frac{1}{2}}\}. \tag{37}$$

The stopping rule is very easy to implement, one simply has to calculate U_n at each stage of sampling and check to see if it falls below cn^2. Moreover, the combined sampling and stopping strategy lead us to the design with an optimal case-control ratio for large n.

4.1. Simulation study for binary exposure

Since all the above results are for large N, we conducted a small scale simulation study to assess the finite sample performance of our design in the binary case. The goal was to estimate β accurately. We started with five cases and five controls in each simulation run. We then generated the data retrospectively at each sampling step, i.e. given that we selected a case ($y_i = 1$) or a control

$(y_i = 0)$ to be sampled at the ith step, we generated the exposure $(x_i|y_i)$ from a Bernoulli distribution with success probability

$$\frac{\exp[(\lambda + \beta y_i)]}{1 + \exp[(\lambda + \beta y_i)]}.$$

We varied the values of c over a grid of 0.05 to 0.0001. For several true values of β and λ and several choices of priors, the stopping time N, the proportion of cases r_N, and the value of cN^2 were noted to validate the results in Theorem 1. Recall that the posterior distribution of $\boldsymbol{\eta}$ was approximated as a bivariate normal distribution

$$N_2((\mathbf{I}_n(\hat{\boldsymbol{\eta}}_n) + \boldsymbol{W}^{-1})^{-1}(\mathbf{I}_n(\hat{\boldsymbol{\eta}}_n)\hat{\boldsymbol{\eta}}_n + \boldsymbol{W}^{-1}\boldsymbol{m}), (\mathbf{I}_n(\hat{\boldsymbol{\eta}}_n) + \boldsymbol{W}^{-1})^{-1}).$$

The approximate posterior mean of β (denoted by $\hat{\beta}_{APM}$, the suffix APM standing for approximate posterior mean) and the MLE of β at the stopping time N (denoted by $\hat{\beta}_{MLE}$) were calculated. Since the Laplace approximation to the posterior mean may not be adequate for small sample sizes, after obtaining the data at stopping time N we estimated the posterior mean by a standard Markov chain Monte Carlo numerical integration technique, using the same prior as used in determining the stopping time. This estimate is denoted by $\hat{\beta}_{MCMC}$. We allowed a burn-in of 20,000 iterations and the final estimate $\hat{\beta}_{MCMC}$ is based on every tenth observation of the last 10,000 runs of five multiple chains, after allowing the burn-in period. Convergence of the chain was assessed by the Gelman and Rubin (1992) diagnostic. As we notice in the results, to the advantage of the practitioner, the Laplace approximation performs adequately in most cases and avoids the computational burden of running an MCMC.

For each configuration of prior parameters, the simulation was repeated 500 times. The results were averaged over the 500 iterations. To establish our convergence results as indicated in Theorem 1, we recorded values of N, r_N and cN^2 for each simulation run. We then calculated the mean and variance of N, r_N and cN^2 over these 500 runs. It was noted that the sequential design strategy indeed leads to the optimal choice of r^* as given in (36), even for moderate sample sizes. The convergence of cN^2 to $g(\lambda, \beta, r^*)$ is also evident.

We now consider the estimation of β, the parameter of interest. Along with the mean of the three estimates of β obtained in 500 runs, we also computed the MSE corresponding to the three estimates of β by averaging the squared deviation of the estimates from the true value over the 500 replications. The performance of the Bayes estimate relative to that of MLE, depends on the choice of the prior parameters, especially for larger values of c (i.e., smaller values of N) and the difference between the two diminishes as c becomes smaller (N becomes larger), i.e. as $G_N - U_N \xrightarrow{P} 0$. For smaller values of N, the Bayes estimate tends to have a smaller MSE than the maximum likelihood estimates, but the gain in terms of MSE largely depends on the selection of priors. For values of β which are small, for example, in Table 1 with $\beta = -1$, the MSE's are somewhat large compared to the effect sizes for small to moderate values of N, when the prior mean (μ_β) is set at 0 with a large prior variability ($\sigma_\beta = 4$). A more informative prior ($\mu_\beta = -1.5$, $\sigma_\beta = 1$) produces much smaller MSE as noted in Table 2. For larger absolute values of β, or larger effect sizes, the relative magnitude of the MSE's for all three estimates are found to be smaller compared to effect sizes (Tables 3, 4). We present our results for larger effect sizes also under two prior choices: (i) a diffuse prior with large variability in Table 3 to reflect the situations where there is lack of precise prior information, and (ii) an informative prior in Table 4 to represent the situations where credible scientific guesses are available. We also include the case of no effect or $\beta = 0$ in the supplementary documentation available online, again with a diffuse prior. In some scenarios, specially for small N, the MCMC estimate of β is better than the estimate obtained by Laplace approximation, $\hat{\beta}_{APM}$, but one must note that the Laplace approximation performs quite comparably with the MCMC estimate and saves enormous computing complexity. In the simulation results, we note that the values of N and the precision of the estimates do depend on the true values that determine the limiting values r^* of r_N and $g(\lambda, \beta, r^*)$ of cN^2, though the basic pattern of the findings remain the same. Mathematica codes for the simulation are available at http://www.stat.ufl.edu/∼mukherjee/research.

Table 1. True values of the parameters: $\lambda = 0.5$, $\beta = -1$, r^*=0.5, $g(r^* = 0.5, \lambda = 0.5, \beta = -1)$=17.021. Prior parameters: $\mu_\lambda = \mu_\beta = 0$, $\sigma_\lambda = 4, \sigma_\beta = 4, \rho = 0.2$. $\hat{\beta}_{APM}$ denotes the posterior mean obtained by using the Laplace approximation, $\hat{\beta_{MCMC}}$ is the exact posterior mean as obtained by implementing the MCMC numerical integration scheme based on the data at stopping time N. The quantities in the parentheses denote the respective MSE's as estimated from the 500 replications.

c	Mean(N) (Var (N))	Mean(r_N) (Var (r_N))	Mean(cN^2) (Var(cN^2))	$\hat{\beta}_{MLE}$ (MSE($\hat{\beta}_{MLE}$))	$\hat{\beta}_{APM}$ (MSE($\hat{\beta}_{APM}$))	$\hat{\beta}_{MCMC}$ (MSE($\hat{\beta}_{MCMC}$))
0.05	19.94 (3.94)	0.4999 (0.002)	20.07 (10.17)	-0.9619 (0.8310)	-0.9130 (0.7398)	-0.9443 (0.7251)
0.02	30.89 (4.19)	0.5002 (0.001)	19.17 (7.11)	-1.0195 (0.5398)	-0.9822 (0.4965)	-1.0001 (0.5002)
0.005	60.18 (5.44)	0.4995 (0.0003)	18.14 (2.07)	-0.9875 (0.2609)	-0.9692 (0.2543)	-0.9704 (0.2402)
0.001	131.57 (9.28)	0.4996 (0.0001)	17.32 (0.66)	-0.9331 (0.1615)	-0.9279 (0.1569)	-0.9301 (0.1623)
0.0001	414.69 (27.21)	0.5005 (0.00005)	17.20 (0.19)	-1.0102 (0.0416)	-1.0100 (0.0407)	-1.0100 (0.0400)

Table 2. True values of the parameters: $\lambda = 0.5$, $\beta = -1$, r^*=0.5, $g(r^* = 0.5, \lambda = 0.5, \beta = -1)$=17.021. Prior parameters: $\mu_\lambda = 0, \mu_\beta = -1.5$, $\sigma_\lambda = 1$, $\sigma_\beta = 1$, $\rho = 0.2$. $\hat{\beta}_{APM}$ denotes the posterior mean obtained by using the Laplace approximation, $\hat{\beta_{MCMC}}$ is the exact posterior mean as obtained by implementing the MCMC numerical integration scheme based on the data at stopping time N. The quantities in the parentheses denote the respective MSE's as estimated from the 500 replications.

c	Mean(N)	Mean(r_N)	Mean(cN^2)	$\hat{\beta}_{MLE}$	$\hat{\beta}_{APM}$	$\hat{\beta}_{MCMC}$
	(Var (N))	(Var (r_N))	(Var(cN^2))	(MSE($\hat{\beta}_{MLE}$))	(MSE($\hat{\beta}_{APM}$))	(MSE($\hat{\beta}_{MCMC}$))
0.05	19.86 (2.02)	0.5032 (0.00165)	19.821 (8.85)	-0.8862 (0.7985)	-1.0164 (0.3196)	-1.0239 (0.3041)
0.02	30.91 (2.38)	0.5019 (0.00093)	19.13 (3.81)	-1.0442 (0.7190)	-1.0719 (0.3696)	-1.0632 (0.3421)
0.005	59.69 (3.85)	0.5023 (0.00028)	17.83 (1.41)	-0.9608 (0.2408)	-1.0015 (0.1668)	-0.9999 (0.1509)
0.001	132.35 (7.87)	0.5000 (0.00014)	17.52 (0.56)	-1.0214 (0.1173)	-1.0350 (0.0995)	-1.0285 (0.1035)
0.0001	415.32 (36.48)	0.4998 (0.00004)	17.25 (0.25)	-1.0415 (0.0546)	-1.0452 (0.0528)	-1.0457 (0.0507)

Table 3. True values of the parameters: $\lambda = 1$, $\beta = -3$, r^*=0.5778, $g(r^* = 0.5778, \lambda = 1, \beta = -3)$=28.531. Prior parameters: $\mu_\lambda = \mu_\beta = 0$, $\sigma_\lambda = \sigma_\beta = 4, \rho = 0.5$. $\hat{\beta}_{APM}$ denotes the posterior mean obtained by using the Laplace approximation, $\hat{\beta_{MCMC}}$ is the exact posterior mean as obtained by implementing the MCMC numerical integration scheme based on the data at stopping time N. The quantities in the parentheses denote the respective MSE's as estimated from the 500 replications.

c	Mean(N)	Mean(r_N)	Mean(cN^2)	$\hat{\beta}_{MLE}$	$\hat{\beta}_{APM}$	$\hat{\beta}_{MCMC}$
	(Var (N))	(Var (r_N))	(Var(cN^2))	(MSE($\hat{\beta}_{MLE}$))	(MSE($\hat{\beta}_{APM}$))	(MSE($\hat{\beta}_{MCMC}$))
0.05	26.52 (32.47)	0.5787 (0.00941)	36.79 (324.10)	-2.899 (1.0621)	-2.672 (0.9313)	-2.713 (0.9112)
0.02	41.01 (52.45)	0.5841 (0.00511)	34.67 (188.23)	-2.949 (0.7357)	-2.802 (0.6799)	-2.815 (0.6524)
0.005	79.07 (90.34)	0.5876 (0.0023)	31.71 (63.86)	-3.001 (0.3980)	-2.921 (0.3693)	-2.919 (0.3710)
0.001	171.49 (195.24)	0.5768 (0.0012)	29.60 (24.68)	-2.999 (0.1784)	-2.965 (0.1732)	-2.973 (0.1741)
0.0001	536.53 (534.51)	0.5784 (0.00033)	28.84 (6.32)	-2.993 (0.0568)	-2.983 (0.0564)	-2.991 (0.0563)

Table 4. True values of the parameters: $\lambda = -1$, $\beta = 4$, r^*=0.676, $g(r^* = 0.676, \lambda = -1\beta = 4)$=48.443. Prior parameters: $\mu_\lambda = 0, \mu_\beta = 5$, $\sigma_\lambda = 2, \sigma_\beta = 1, \rho = 0.2$. $\hat{\beta}_{APM}$ denotes the posterior mean obtained by using the Laplace approximation, $\beta_{M\hat{C}MC}$ is the exact posterior mean as obtained by implementing the MCMC numerical integration scheme based on the data at stopping time N. The quantities in the parentheses denote the respective MSE's as estimated from the 500 replications.

c	Mean(N)	Mean(r_N)	Mean(cN^2)	$\hat{\beta}_{MLE}$	$\hat{\beta}_{APM}$	$\hat{\beta}_{MCMC}$
	(Var (N))	(Var (r_N))	(Var(cN^2))	(MSE($\hat{\beta}_{MLE}$))	(MSE($\hat{\beta}_{APM}$))	(MSE($\hat{\beta}_{MCMC}$))
0.05	39.44	0.7132	94.31	4.104	4.238	4.329
	(333.96)	(0.0200)	(15562.20)	(1.003)	(0.4658)	(0.4322)
0.02	56.11	0.6977	67.68	4.0763	4.1859	4.2196
	(237.84)	(0.0085)	(1497.24)	(1.0567)	(0.6189)	(0.6207)
0.005	100.83	0.6715	52.48	3.9173	4.0144	4.0007
	(333.48)	(0.0039)	(389.31)	(0.4189)	(0.3098)	(0.2887)
0.001	227.26	0.6724	50.12	4.0271	4.0687	4.0183
	(488.01)	(0.0015)	(103.17)	(0.1443)	(0.1288)	(0.1345)
0.0001	694.46	0.6775	48.37	3.9321	3.9479	3.9529
	(1455.36)	(0.0003)	(28.52)	(0.0553)	(0.0512)	(0.0527)

5. Extension to Group Sequential Design and Matched Case-Control Study

The sampling scheme and the stopping rule, as described above, is based on a purely sequential strategy where one stops at each stage to choose a case or a control. Practitioners often prefer interim analysis based on a batch sequential strategy, especially when the allowable sample size is fairly large and it may take many steps of sampling before one stops. Group sequential methods in a fully decision theoretic Bayesian framework have been limited because of the great analytical and computational complexity in implementing solutions via backward induction (Degroot (1970)). The main literature has been restricted to applications related to clinical trials with simple model settings, like one sided tests with binary outcomes (Lewis and Berry (1994)), and few (typically two to five) backward steps. Carlin, Kadane and Gelfand (1998) propose a forward sampling algorithm that eases the computational burden of backward induction, offering the possibility of many interim looks. For a detailed review of group-sequential methods, see Jennison and Turnbull (2000).

In this section we briefly indicate how our proposed methods could be extended to the group sequential framework. The extension of the sampling scheme, as indicated in Chen (2000), is briefly described as follows.

- Let m_k denote the size of the kth batch. We consider (i) the m_k's are fixed by the practitioner, the more common scenario, and (ii) the m_k's are to be determined at each step.
- Let m_{1k} denote the number of cases and m_{0k} denote the number of controls to be chosen in batch k, with $m_k = m_{0k} + m_{1k}$.
- Let $\hat{r}_k^*$ denote the solution to $\hat{D}(r) = 0$ with the current estimate $\hat{D}$ at stage k.
- Let $n_k = \sum_{j=1}^k m_j$ be the total sample size after batch k is drawn, with $n_{1k} = \sum_{j=1}^k m_{1j}$ and $n_{0k} = \sum_{j=1}^k m_{0j}$ the total number of cases and controls after stage k, $n_k = n_{0k} + n_{1k}$.

Case (i). If the m_k's are predetermined by the practitioner, one simply needs to decide on how to allocate the cases and the controls at each stage. Suppose one has completed sampling the $(k-1)$th batch and is about to select the kth batch.

The proportion of cases after stage k, if one chooses m_{1k} cases, is $r_k = (n_{1k-1} + m_{1k})/(n_{k-1} + m_k)$. We recommend the following choice for m_{1k}:

$$m_{1k} = \max\{\min(\hat{r}_{k-1}^*(n_{k-1} + m_k) - n_{1k-1}, m_k), 0\}. \tag{38}$$

This choice minimizes the absolute departure of r_k from r_{k-1}^*. After the kth stage,

$$r_k = \max\left\{\frac{n_{1k-1}}{n_{k-1} + m_k}, \min\left(\hat{r}_{k-1}^*, \frac{n_{1k-1} + m_k}{n_{k-1} + m_k}\right)\right\}. \tag{39}$$

If m_k is small, instead of solving $\hat{D}(r) = 0$ at each stage one could go by the sign of $\hat{D}(r_{nk-1})$ and choose all cases (controls) if the sign is negative (positive), exactly as in the purely sequential case.

Case (ii). The batch sizes m_k are unknown, so one has to decide on both m_k and m_{1k}. One may choose m_k and m_{1k} to satisfy

$$\frac{n_{1k-1} + m_{1k}}{n_{k-1} + m_k} = \hat{r}_{k-1}^*. \tag{40}$$

Since there are infinitely many solutions to the above equation, we suggest the following strategy.

- If $n_{1k-1}/n_{k-1} < \hat{r}_{k-1}^*$, choose no controls and $(n_{k-1}\hat{r}_{k-1}^* - n_{1k-1})/(1 - \hat{r}_{k-1}^*)$ cases.
- If $n_{1k-1}/n_{k-1} > \hat{r}_{k-1}^*$, choose no cases and $(-n_{k-1}\hat{r}_{k-1}^* + n_{1k-1})/\hat{r}_{k-1}^*)$ controls.
- If $n_{1k-1}/n_{k-1} = \hat{r}_{k-1}^*$, choose an equal number of cases and controls.

When the evaluated expression for the number of cases (controls) does not result into integers, one could approximate the number of cases by the greatest (smallest) integer contained in (exceeding) the evaluated expression.

The Stopping Rule. The APO rule is a straightforward generalization of the purely sequential case. Stop at the Nth batch, where

$$N = \inf\{J \geq 1 : \sum_{k=1}^{J} m_k = n_J; n_J \geq (\frac{G_{n_J}}{c})^{\frac{1}{2}}\}. \tag{41}$$

Proofs for asymptotic properties of the sampling design in the group sequential case can be obtained in a similar way, but necessitate complex computations that are omitted in the current paper.

Remark 7. The $1 : m_{0k}$ matched case-control analysis can be viewed as a special case of (40) with $m_{1k} \equiv 1$ and $m_k = 1 + m_{0k}$. In the matched design, one chooses m_{0k} satisfying (40). This is the strategy one might implement in a synthetic case-control design, or a case-control within a cohort study where a "desirable" number of controls are selected when a case occurs. More typically, for a case-control study, all available cases (say n_k) over a time interval are selected and the number of controls (m_k) is chosen after the cases are recruited. The strategy for choosing m_k, as described in (40), could be adopted in this setting as well, with the given value of n_k. We would like to point out that the appropriate likelihood for such sampling designs is a weighted conditional logistic likelihood, and using a likelihood such as the one used in this paper will only give one a rough idea regarding the approximate number of controls to be selected at each time point. The asymptotics using the exact conditional logistic likelihood and a similar sequential treatment appears to be much more complex, and remains beyond the scope of this paper.

Remark 8. As one referee has pointed out, case-control studies differ from randomized clinical trials in the sense that in a case-control study, the investigators do not provide an intervention; thus, the need for interim looks at the data (for safety monitoring purposes) is far smaller. So the classical set-up of group sequential framework is less appealing in a case-control study. The current paper is indeed quite different from the popular sequential clinical trial designs in its objective and implementation. Here we address the problem of sequentially determining the optimal case-control ratio where the prespecified goal is precision in estimating β. If interim evaluation of the optimal case-control ratio is hard, one could introduce a cost for performing interim analysis in the loss function and address the problem in a Bayesian decision theoretic framework. Since the method is computationally very simple, and sampling or assaying a frozen serum sample as needed in the Nurse's Health Study could be quite expensive, the proposed sequential method could save considerable resources. The decision theoretic approach could also motivate new design strategies in the context of similar practical problems related to a case-control sampling design where determining the sampling ratio is of critical importance (Morgenstern and Winn (1983)).

Acknowledgements

This article is devoted to the memory of the late Professor Ching-Zong Wei. Ghosh's research was partially supported by NSF Grant SES-0317585. Mukherjee's research was partially supported by NSA Young Investigator Grant H98230-06-1-0033. The authors wish to thank two referees and an associate editor for their constructive suggestions.

References

Baron, M. (2002). Bayes and asymptotically pointwise optimal stopping rules for the detection of influenza epidemics, In *Case Studies in Bayesian Statistics* **6** (Edited by C. Gastonis, R. E. Kass, A. Cariquirry, A. Gelman, D. Higdon, D. K. Pauler and I. Verdinelli), Springer-Verlag, New York.

Bickel, P. and Yahav, J. (1967). Asymptotically Pointwise Optimal procedures in sequential analysis. *Proceedings of* 5*th. Berkeley Symposium* **VI**, 401-413.

Bickel, P. and Yahav, J. (1968). Asymptotically optimal Bayes and minimax procedures in sequential estimation. *Ann. Math. Statist.* **39**, 442-456.

Bickel, P. and Yahav, J. (1969). An A.P.O. rule in sequential estimation with quadratic loss. *Ann. Math. Statist.* **40**, 417-426.

Carlin, B. P., Kadane, J. B. and Gelfand, A. E. (1998). Approaches for optimal sequential decision analysis in clinical trials, *Biometrics* **54**, 964-975.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian designs applied to logistic regression experiments. *J. Statist. Plann. Inference* **21**, 191-208.

Chen, K. (2000). Optimal sequential design for case-control studies. *Ann. Statist.* **28**, 1452-1471.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* **10**, 417-451.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113-120.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion), *J. Roy. Statist. Soc. Ser. B, Methodological* **30**, 248-275.

Dalal, S. R. and Mallows, C. L. (1988). When should one stop testing software? *J. Amer. Statist. Asssoc.* **83**, 872-879.

Degroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

Fakhre-Zakeri, I. and Slud, E. (1996). Optimal stopping of sequential size-dependent search. *Ann. Statist.* **24**, 2215-2232.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-472.

Gleser, L. J. and Kunte, S. (1976). On asymptotically optimal sequential Bayes interval estimation procedures. *Ann. Statist.* **4**, 685-711.

Hoekstra, R. M. (1989). Asymptotically pointwise optimal stopping rules in multiparameter estimation. Ph.D. Dissertation, Department of Statistics, University of Florida.

Jennison C. and Turnbull B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall, CRC Press, Boca Raton.

Lewis, R. J. and Berry, D. A. (1994). Group sequential clinical trials: A classical evaluation of Bayesian decision theoretic designs. *J. Amer. Statist. Assoc.* **89**, 1528-1534.

Lui, K.-J. (1990). Sample size determination for case-control studies: The influence of the joint distribution of exposure and confounder. *Statist. Medicine* **9**, 1485-1493.

Lui, K-J. (1993). Sample size determination for multiple continuous risk factors in case-control studies. *Biometrics*, **49**, 873-876.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719-748.

Morgenstern, H. and Winn, D. M. (1983), A method for determining the sampling ratio in epidemiologic studies, *Statist. Medicine* **2**, 387-396.

Munoz, A. and Rosner, B. (1994). Power and sample size for a collection of 2 × 2 tables. *Biometrics* **40**, 995-1004.

Nam, J.-M. (1992). Sample size determination for case-control studies and the comparison of stratified and unstratified analyses. *Biometrics* **48**, 389-395.

Nam, J-M. (1997). Sample size determination for designing a strata-matched case-control study to detect multiple risk factors. *Biometrical J.* **39**, 441-454.

Nam, J-M., and Fears, T. R. (1990). Optimum allocation of samples in strata-matching case-control studies when cost per sample differs from stratum to stratum. *Statist. Medicine* **9**, 1475-1483.

Nam, J-M , and Fears, T. R. (1992a). Optimum sample size determination in stratified case-control studies with cost considerations. *Statist. Medicine* **11**, 547-556.

Nam, J.-M. and Fears, T. R. (1992b). Control sample size when cases are given in constant ratio stratum-matched case-control studies, *Statist. Medicine* **11**, 1759-1766.

O'Neill, R. T. (1998). Case-control study, sequential. In *Encyclopedia of Biostatistics* **1** (Edited by P. Armitage and T. Colton), 528-532. Wiley, New York.

O'Neill, R. T. and Anello, C. (1978). Case-control studies: A sequential approach. *Amer. J. Epidemiology* **120**, 145-153.

Pasternack, B. S. and Shore, R. E. (1981). Sample sizes for individually matched case-control studies. *Amer. J. Epidemiology* **115**, 778-784.

Prentice, R. L. (1986). On the design of synthetic case-control studies. *Biometrics* **42**, 301-310.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Rehalia, M. E. H. (1984). Asymptotic sequential analysis in the A.P.O. rule performance. *Sequential Anal.* **3**, 155-174.

Schlesselman, J. J. (1974). Sample size requirements in cohort and case-control studies of disease. *Amer. J. Epidemiology* **99**, 381-384.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer Series in Statistics, Springer - Verlag, New York.

Stampfer, M. J., Willet, W. C., Coldits, G., Roser, B., Speizer, F. and Henekens, C. (1985). A prospective study of postmenopausal estrogen therapy and coronary heart disease. *New England J. Medicine* **313**, 1044-1049.

Vessey, P. M. and Doll, D. R. (1968). Investigation of relation between oral contraceptive and thromboembolic disease. *British Medical J.* **2**, 199-205.

Whittemore, A. S. (1981). The efficiency of synthetic retrospective studies. *Biom. J.* **23**, 73-78.

Woodroofe, M. (1981). A.P.O. rules are asymptotically non-deficient for estimation with squared error loss. *Z. Wahrsch. Verw. Gebiete* **58**, 331-341.

Woolson, R. F., Bean, J. A. and Rojas, P. B. (1986). Sample size for case-control studies using Cochran's statistic. *Biometrics* **42**, 927-932.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *J. Amer. Statist. Assoc.* **80**, 974-984.

Zelen, M. and Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statist. Medicine* **5**, 261-269.

Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.

E-mail: ghoshm@stat.ufl.edu

Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.

E-mail: mukherjee@stat.ufl.edu

(Received July 2005; accepted April 2006)