Biometrika (2004), **91**, 1, *pp.* 15–25
© 2004 Biometrika Trust
Printed in Great Britain

Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies

By SHAUN R. SEAMAN

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, U.K.
seaman@mrc-bsu.cam.ac.uk

and SYLVIA RICHARDSON

Department of Epidemiology and Public Health, Imperial College, London W2 1PG, U.K.
sylvia.richardson@ic.ac.uk

Summary

The natural likelihood to use for a case-control study is a 'retrospective' likelihood, i.e. a likelihood based on the probability of exposure given disease status. Prentice & Pyke (1979) showed that, when a logistic regression form is assumed for the probability of disease given exposure, the maximum likelihood estimators and asymptotic covariance matrix of the log odds ratios obtained from the retrospective likelihood are the same as those obtained from the 'prospective' likelihood, i.e. that based on probability of disease given exposure. We prove a similar result for the posterior distribution of the log odds ratios in a Bayesian analysis. This means that the Bayesian analysis of case-control studies may be done using a relatively simple model, the logistic regression model, which treats data as though generated prospectively and which does not involve nuisance parameters for the exposure distribution.

Some key words: Bayesian inference; Case-control study; Dirichlet distribution; Markov chain Monte Carlo; Retrospective likelihood.

1. Introduction

In a case-control study, subjects are recruited according to their disease status and then their past exposure is determined. Thus, the natural likelihood is the 'retrospective' likelihood, i.e. that based on probability of exposure given disease, which involves many nuisance parameters. For maximum likelihood estimation, Prentice & Pyke (1979) showed that a likelihood based on the probability of disease given exposure, the 'prospective' likelihood, gives identical inferences; that is, they showed that, when a logistic regression form is assumed for the probability of disease given exposure, both the prospective and retrospective likelihoods lead to the same maximum likelihood estimators of exposure effects and asymptotic covariance matrix. The prospective likelihood has many fewer nuisance parameters and is computationally much easier to handle than the retrospective likelihood. The result of Prentice & Pyke meant that case-control studies could be analysed using logistic regression, as though disease were the dependent variable and exposure the independent variable. No such result has hitherto been demonstrated for Bayesian inference.

Models for the Bayesian analysis of particular types of case-control study have been described: Zelen & Parker (1986), Nurminen & Mutanen (1987), Marshall (1988) and Ashby et al. (1993) deal with a single binary exposure; Raghunathan (1994) covers a single binary exposure with confounding factors; Müller & Roeder (1997) treat a bivariate continuous exposure; Müller et al. (1999) allow any number of continuous or binary exposures; and Seaman & Richardson (2001) deal with categorical exposure variables. Apart from those for a single binary exposure, these models are complicated, because they use the retrospective likelihood. They have been fitted using Markov chain Monte Carlo methods and each has required problem-specific algorithms and computer code.

Gustafson et al. (2002), treating the problem of measurement error in exposure, allow both discrete and continuous exposures and use a Dirichlet prior that only places support on a grid of possible exposure values. In the absence of measurement error, they suggest that the grid points should simply be the set of unique observed exposure values, pointing out that this device of pretending that continuous observations arose from a discrete distribution with support equal to the observed values is also used by the Bayesian bootstrap (Rubin, 1981). Their algorithm generates posterior samples from the prospective model and then uses importance weighting to estimate the posterior distribution from the retrospective model. They find that, when sufficient quantities of data are observed, the importance-weighting step makes little difference to the results, meaning that the prospective and retrospective models give similar, although not identical, results.

In this paper we prove that a Bayesian analysis that uses the prospective likelihood, and assumes a uniform prior distribution for the log odds that an individual with baseline exposure is diseased, is exactly equivalent to an analysis that uses the retrospective likelihood and assumes a Dirichlet prior distribution for the exposure probabilities in the control group. This means that the Bayesian analysis of case-control studies may, like the classical frequentist analysis, be carried out using a prospective model, thus significantly reducing its complexity.

The Bayesian framework, combined with the Markov chain Monte Carlo technology, offers many possibilities for the flexible modelling of data (Gilks et al., 1996). However, until now, most models created have been for prospective data, such as cohort data. This paper enables the advantages of the Bayesian approach to be exploited easily for modelling case-control data. Possibilities include random-effects models, variable selection (George & McCulloch, 1996), partition models (Consonni & Veronese, 1995; Seaman et al., 2002) and, of course, use of prior information about odds ratios.

The proof that follows is based on the multinomial-Poisson transformation (Baker, 1994). We first show how the transformation may be employed in the classical framework to demonstrate equality of profile likelihoods from prospective and retrospective models. This proof is in §2 and is followed, in §3, by a proof of equality of posterior distributions in the Bayesian framework. Section 4 treats a particular limiting case of the prior distribution for exposure. In §5 we discuss limitations of the Bayesian equivalence result of this paper, most notably that it cannot be used when the exposure distribution has to be modelled. Section 6 contains an example of how a previously published model, which was fitted using the retrospective likelihood, can be much more easily fitted using the prospective likelihood. We end with a discussion.

2. Equivalence for the classical analysis

We have outlined Prentice & Pyke's (1979) approach in §1. A different approach was followed by Roeder et al. (1996). Their aim was to extend Prentice & Pyke's equivalence result to the case where covariates are measured with error. They proved that the prospective and retrospective models generate the same profile likelihood for the log odds ratio, a proof which also applies when there is no measurement error. However, if one or more covariates are continuous, the retrospective model becomes semiparametric and the usual theoretical justification for using the profile likelihood to calculate maximum likelihood estimates and confidence intervals fails. Later, Murphy & van der Vaart (2000) provided the justification, demonstrating consistency and asymptotic normality of the maximum likelihood estimates from the profile likelihood. Their proof, however, assumes that some measurement error is present.

We now present an alternative proof of the equality of profile likelihoods in the absence of measurement error. The purpose of this is to introduce the multinomial-Poisson transformation, which will be used in §3 to prove equality of posterior distributions in the Bayesian framework. Although the following proof would also work for continuous covariates, we limit ourselves to discrete exposure variables.

Suppose that a study is conducted in which a number of subjects are recruited from the population. Some are diseased and the others are undiseased. A vector X, possibly of length one, of discrete exposure variables is observed for each subject. Let the support of X be $\{z_1, \ldots, z_J\}$, and let Y_{0j} and Y_{1j} be respectively the numbers of undiseased and diseased subjects having $X = z_j$, for $j = 1, \ldots, J$. Suppose that the log odds ratio of disease associated with $X = x$ is $\delta^{\mathrm{T}} x$, where δ is a vector of unknown parameters.

If the study is a case-control study, the natural likelihood is the 'retrospective' likelihood,

$$L_{\mathrm{MR}}(\beta, \delta; y) = \prod_{d=0}^{1} \prod_{j=1}^{J} \left\{ \frac{\beta_j \exp(d\delta^{\mathrm{T}} z_j)}{\sum_{k=1}^{J} \beta_k \exp(d\delta^{\mathrm{T}} z_k)} \right\}^{y_{dj}}.$$

The probability that a control has exposure z_j is $\beta_j / \sum_{k=1}^{J} \beta_k$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_J)$. For identifiability, we assume that $\beta_1 = 1$. Observe that L_{MR} is a product of two independent multinomial likelihoods.

If the study is a cohort study, the natural likelihood is the 'prospective' likelihood,

$$L_{\mathrm{MP}}(\alpha, \delta; y) = \prod_{j=1}^{J} \prod_{d=0}^{1} \left\{ \frac{\alpha^d \exp(d\delta^{\mathrm{T}} z_j)}{\sum_{k=0}^{1} \alpha^k \exp(d\delta^{\mathrm{T}} z_j)} \right\}^{y_{dj}}.$$

Parameter α is the baseline odds of disease, i.e. the odds of disease when exposure is zero. Observe that L_{MP} is a product of J independent binomial likelihoods.

Theorem 1. *The profile likelihood of δ obtained by maximising L_{MR} with respect to β is the same as the profile likelihood of δ obtained by maximising L_{MP} with respect to α.*

Proof. Suppose that random variables Y_{dj} $(d = 0, 1;\ j = 1, \ldots, J)$ are independently distributed as $Y_{dj} \sim \mathrm{Po}(\lambda_{dj})$, where

$$\log \lambda_{dj} = \log \mu + d \log \alpha + \log \beta_j + d\delta^{\mathrm{T}} z_j,$$

with $\beta_1 = 1$ for identifiability. The likelihood for $(\mu, \alpha, \beta, \delta)$ is

$$L_{\mathrm{Po}}(\mu, \alpha, \beta, \delta; y) = \prod_{d=0}^{1} \prod_{j=1}^{J} (\lambda_{dj})^{y_{dj}} \exp(-\lambda_{dj}).$$

Baker (1994) points out that L_{MR} is the 'profile' likelihood for (β, δ) after maximisation of L_{Po} with respect to (μ, α). This is the so-called 'multinomial-Poisson transformation'.

Similarly, the 'profile' likelihood for (α, δ) after maximisation of L_{Po} with respect to (μ, β) is L_{MP}. Since the order of maximisation is immaterial, it follows that the 'profile' likelihood for δ after maximising over nuisance parameters is identical whether we start from L_{Po}, L_{MR} or L_{MP}. □

The natural likelihood for a case-control study is L_{MR}, but maximisation of L_{MP} with respect to α and δ is easier: it can be done using any of the logistic regression routines built into most statistical software.

3. Equivalence for the Bayesian analysis

The proof of equivalence in the Bayesian framework resembles that of the preceding section, but requires integration, rather than maximisation, over nuisance parameters. Crucially, it also involves prior distributions for nuisance parameters.

Theorem 2. *Suppose that random variables Y_{dj} $(d = 0, 1;\ j = 1, \ldots, J)$ are independently distributed as $Y_{dj} \sim \mathrm{Po}\,(\lambda_{dj})$, where*

$$\log \lambda_{dj} = d \log \alpha + \log \beta_j + d\delta^{\mathrm{T}} z_j.$$

Assume independent improper priors, $p(\alpha) \propto \alpha^{-1}$ and $p(\beta_j) \propto \beta_j^{a_j - 1}$, for α and β, and a prior, $p(\delta)$, for δ that is independent of α and β and satisfies the condition that, for some q and r such that $y_{0q} \geqslant 1$ and $y_{0r} \geqslant 1$, $E(z_q^{\mathrm{T}}\delta)$ and $E(z_r^{\mathrm{T}}\delta)$ exist and are finite. Let $y_{+j} = y_{0j} + y_{1j}$ and $y_{d+} = \sum_{j=1}^{J} y_{dj}$. Then the following two statements hold.

(i) *The posterior density of (ω, δ), where $\omega = \log \alpha$, is*

$$p(\omega, \delta | y) \propto p(\delta) \prod_{j=1}^{J} \frac{\{\exp(\omega + \delta^{\mathrm{T}} z_j)\}^{y_{1j}}}{\{1 + \exp(\omega + \delta^{\mathrm{T}} z_j)\}^{y_{+j} + a_j}}. \qquad (1)$$

(ii) *The posterior density of (θ, δ), where $\theta_j = \beta_j / \sum_{k=1}^{J} \beta_k$ and $\theta = (\theta_1, \ldots, \theta_J)$, is*

$$p(\theta, \delta | y) \propto p(\delta) \prod_{j=1}^{J} \theta_j^{a_j - 1} \prod_{d=0}^{1} \left[\frac{\prod_{j=1}^{J} \{\theta_j \exp(d\delta^{\mathrm{T}} z_j)\}^{y_{dj}}}{\{\sum_{j=1}^{J} \theta_j \exp(d\delta^{\mathrm{T}} z_j)\}^{y_{d+}}} \right]. \qquad (2)$$

The proportionality is proportionality up to a constant.

Proof. (i) The posterior density of (α, β, δ) is

$$p(\alpha, \beta, \delta | y) \propto p(\delta) \frac{1}{\alpha} \prod_{j=1}^{J} \beta_j^{a_j - 1} \prod_{d=0}^{1} \prod_{j=1}^{J} (\lambda_{dj})^{y_{dj}} \exp(-\lambda_{dj}). \qquad (3)$$

It is not immediately obvious that this is a proper distribution under the stated conditions for $p(\delta)$, but this is proved in the Appendix. Integrating with respect to β and then performing a transformation of variable from α to ω yields density (1).

(ii) Starting from equation (3), transform from β to (θ, ψ), where $\psi = \sum_{j=1}^{J} \beta_j$, and then integrate out α followed by ψ. The Jacobian of the transformation from β to (θ, ψ) is (Devroye, 1986, p. 594)

$$\left| \frac{\partial(\beta_1, \ldots, \beta_J)}{\partial(\theta_1, \ldots, \theta_{J-1}, \psi)} \right| = \psi^{J-1}.$$

Hence,

$$p(\alpha, \theta, \psi, \delta|y) \propto p(\delta)\alpha^{y_{1+}-1}\psi^{y_{++}+\alpha_{+}-1}\prod_{j=1}^{J}\theta_j^{y_{+j}+a_j-1}\exp\left(\sum_{j=1}^{J}\delta^{\mathrm{T}}z_j y_{1j}\right)$$

$$\times\exp\left(-\psi\left[\sum_{j=1}^{J}\theta_j\{1+\alpha\exp(\delta^{\mathrm{T}}z_j)\}\right]\right),$$

where $y_{++}=\sum_{d=0}^{1}\sum_{j=1}^{J}y_{dj}$ and $a_{+}=\sum_{j=1}^{J}a_j$. Integrating over α and then ψ, we obtain density (2). □

Note that the condition that $E(z_q^{\mathrm{T}}\delta)$ and $E(z_r^{\mathrm{T}}\delta)$ exist and are finite is automatically satisfied if the prior $p(\delta)$ is such that $E(\delta)$ exists and is finite.

COROLLARY 1. *The marginal posterior densities of δ obtained from joint densities* (1) *and* (2) *are the same.*

Proof. The order in which one integrates out the joint posterior, $p(\alpha, \beta, \delta|y)$, of the Poisson model in order to obtain the marginal posterior density of δ makes no difference. Thus one may integrate equation (1) with respect to ω or equation (2) with respect to θ, and still obtain the same $p(\delta|y)$. □

We now interpret Corollary 1 in the context of a case-control study. Suppose a vector, X, of exposure variables is observed for each subject in a case-control study. If all the exposure variables are discrete, let the support of X be $\{z_1, \ldots, z_J\}$. If some or all of them are continuous, we either discretise them or follow the Bayesian bootstrap and Gustafson et al. (2002) in pretending that only exposure values actually observed may be observed, so that $\{z_1, \ldots, z_J\}$ are the exposure values observed in the study. Let y_{0j} and y_{1j} be the numbers of controls and cases respectively with $X=z_j$.

Equation (1) corresponds to a prospective model for the data, in which ω is the baseline log odds of disease. If all a_j's are integers, $p(\omega, \delta|y)$ is the posterior density of (ω, δ) from the model

$$Y_{1j}\sim \mathrm{Bi}(Y_{+j}+a_j, p_j)\quad (j=1, \ldots, J),$$

$$\log\left(\frac{p_j}{1-p_j}\right)=\omega+\delta^{\mathrm{T}}z_j$$

with priors $p(\delta)$ and the improper $p(\omega)\propto 1$. This is the prospective model with a_j extra undiseased individuals with exposure z_j having been observed.

Equation (2) corresponds to a retrospective model for the data, in which θ_j is the probability that a control has exposure z_j:

$$(Y_{d1}, Y_{d2}, \ldots, Y_{dJ})\sim \mathrm{Mu}(Y_{d+}; p_{d1}, p_{d2}, \ldots, p_{dJ})\quad (d=0, 1),$$

$$p_{dj}=\frac{\theta_j\exp(d\delta^{\mathrm{T}}z_j)}{\sum_{k=1}^{J}\theta_k\exp(d\delta^{\mathrm{T}}z_k)}\quad (j=1, \ldots, J),$$

with priors $p(\delta)$ and $\theta\sim \mathrm{Dir}(a_1, \ldots, a_{\mathrm{J}})$.

Thus, Corollary 1 states that one may fit either the prospective or retrospective model and obtain the same posterior marginal distribution for the parameter of interest, δ. The prospective and retrospective models may be said to be equivalent. Whereas the retrospective model involves J nuisance parameters, the prospective model involves only one. In §6, we show that the prospective model can easily be fitted.

4. A limiting prior distribution for exposure

The Dirichlet prior is conjugate for the multinomial likelihood: if

$$W=(W_1, \ldots, W_J) \sim \mathrm{Mu}(\theta_1, \ldots, \theta_J), \quad \theta=(\theta_1, \ldots, \theta_J) \sim \mathrm{Dir}(a_1, \ldots, a_J),$$

then $\theta | W \sim \mathrm{Dir}(a_1+W_1, \ldots, a_J+W_J)$. Thus, a_j may be interpreted as the number of counts in the jth category observed prior to the study. Furthermore, if $\theta \sim \mathrm{Dir}(a, \ldots, a)$, then $E(\theta_j)=J^{-1}$ and $\mathrm{var}(\theta_j)=(J-1)/\{J^2(Ja+1)\}$. The variance of θ_j therefore increases as a decreases. It is for these reasons that the $\mathrm{Dir}(a, \ldots, a)$ distribution in the limit as a tends to zero is often thought of as representing a noninformative prior. In the context of the retrospective model, a_j may be interpreted as the number of undiseased individuals with exposure z_j observed prior to the study.

For all $a>0$, $p(\alpha, \beta, \delta | y)$ is proper. Also, if for all $j=1, \ldots, J$ at least one subject is observed with $X=z_j$, Theorem 2 still applies when $a=0$. However, if for any of the support points, z_j, of the exposure distribution no subject is observed, the posterior density $p(\alpha, \beta, \delta | y)$ is improper when $a=0$. It is therefore necessary to examine the equivalence relation as a tends to zero.

THEOREM 3. *As a tends to zero, the log odds ratio parameters, δ, from the retrospective model converge in distribution to the log odds ratio parameters from the 'limiting' prospective model, given by*

$$Y_{1j} \sim \mathrm{Bi}(Y_{+j}, p_j) \quad (j=1, \ldots, J),$$

$$\log\left(\frac{p_j}{1-p_j}\right)=\omega+\delta^{\mathrm{T}} z_j,$$

$$\delta \sim p(\delta), \quad p(\omega) \propto 1.$$

Proof. Let ω_n and δ_n be sequences of random variables with joint probability density $q_n(\omega, \delta)$, where

$$q_n(\omega, \delta)=r(\delta) \prod_{j=1}^{J} \frac{\{\exp(\omega+\delta^{\mathrm{T}} z_j)\}^{y_{1j}}}{\{1+\exp(\omega+\delta^{\mathrm{T}} z_j)\}^{y_{+j}+1/n}}$$
$$\times\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(v) \prod_{j=1}^{J} \frac{\{\exp(u+v^{\mathrm{T}} z_j)\}^{y_{1j}}}{\{1+\exp(u+v^{\mathrm{T}} z_j)\}^{y_{+j}+1/n}} du\, dv\right]^{-1}$$

and $r(.)$ is any proper probability distribution. Using Lebesgue's dominated convergence theorem (Billingsley, 1986, p. 213), we have

$$\lim_{n \to \infty} q_n(\omega, \delta)=r(\delta) \prod_{j=1}^{J} \frac{\{\exp(\omega+\delta^{\mathrm{T}} z_j)\}^{y_{1j}}}{\{1+\exp(\omega+\delta^{\mathrm{T}} z_j)\}^{y_{+j}}}$$
$$\times\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(v) \prod_{j=1}^{J} \frac{\exp\{(u+v^{\mathrm{T}} z_j)\}^{y_{1j}}}{\{1+\exp(u+v^{\mathrm{T}} z_j)\}^{y_{+j}}} du\, dv\right]^{-1}. \qquad (4)$$

It now follows from Scheffé's Theorem (Billingsley, 1986, p. 218) that, as n tends to ∞, δ_n converges in distribution to the random variable δ, where the probability density function of δ is given by the integral of the right-hand side of equation (4) with respect to ω.

Therefore, if a proper $\mathrm{Dir}(a, \ldots, a)$ prior, with $a>0$, is used, then, as a becomes closer and closer to zero, the posterior density $p(\delta | y)$ from the limiting prospective model becomes an increasingly exact approximation of $p(\delta | y)$ from the retrospective model. □

Note that the limiting prospective model does not depend on the support $(z_1, \ldots, z_J)$. This makes it particularly suitable when one is modelling continuous exposure variables by using the pretence that only exposure values that have been observed can be observed; see §3. Even if one allowed other, unobserved, exposure values, the limiting prospective model would remain the same.

5. Limitations of the equivalence result

We have shown that the retrospective model is equivalent to the simpler prospective model if a Dirichlet prior is assumed for exposure in controls. Other authors working with retrospective models for case-control studies have assumed other prior distributions for this exposure distribution. For example, Müller & Roeder (1997) used a mixture of normal distributions with Dirichlet process prior on the mixing measure. We have not shown any equivalence when any prior other than the Dirichlet is used. The fact that the prospective model does not contain parameters for the exposure distribution is both an advantage and a potential disadvantage. The prospective model is easier to fit than the retrospective model, but it cannot be used when one wishes to model the exposure distribution as well as the odds ratios. This might be, for example, to allow for exposure-measurement error or missing data on exposures. Where the exposure distribution must be modelled, a retrospective model is required. Examples of such models are given by Seaman & Richardson (2001), Müller & Roeder (1997) and Gustafson et al. (2002). Like us, Gustafson et al. used a Dirichlet prior and found approximate, although not exact, equivalence of retrospective and prospective models. The reason for the lack of equivalence in their approach, even when measurement error is absent, is that their Dirichlet prior describes the marginal exposure distribution in the whole case-control sample, rather than in controls.

This paper has dealt with unstratified case-control studies. The equivalence of prospective and retrospective models does extend to stratified studies: it suffices to have separate ω parameters for each stratum and for these to be a priori independent. However, one needs to be cautious when dealing with small stratum sizes, as the choice of prior distribution for these parameters may become very influential upon the posterior distribution of the parameters of interest. In particular, we have found, proof omitted, that the $\text{Dir}(0, \ldots, 0)$ prior is certainly not noninformative in the extreme case of an individually-matched case-control study, with a single binary exposure. A consistent estimator of the log odds ratio in this case is the conditional maximum likelihood estimator, $\hat{\delta}$, the ratio of the number of discordant pairs in which the case is exposed to the number of discordant pairs in which the control is exposed. We determined that the mode of the posterior distribution of δ, the log odds ratio, is approximately $\exp(\hat{\delta})$, irrespective of the number of case-control pairs. Thus, just as in the classical frequentist framework, where ordinary logistic regression yields biased log odds ratio estimates when strata are small (Pike et al., 1980), the same is true of the Bayesian model with $\text{Dir}(0, \ldots, 0)$ prior. Simulation studies suggested that a stratum size of 10 subjects or fewer is insufficient to make the improper Dirichlet prior noninformative, but that 40 is probably enough.

6. Illustrative example

Seaman & Richardson (2001) have shown how the retrospective model may be fitted using Markov chain Monte Carlo techniques. Although possible, the procedure is demanding.

First, it was necessary to develop Metropolis–Hastings algorithms and to write specific computer code for executing them. Secondly, the mixing of the Markov chains produced by these algorithms needs careful monitoring, because of the need to update both odds-ratio and exposure-distribution parameters, parameters which are heavily correlated. The slow mixing meant long chains had to be generated. In this section we show how the equivalent prospective model may be fitted much more easily and quickly using WinBUGS (Spiegelhalter et al., 1999).

The example used by Seaman & Richardson (2001) was a French case-control study investigating the effect of genotype on risk of lung cancer. They modelled two exposure variables: the CYP2D6 genotype, a six-level factor, and tobacco smoking, coded as low, medium or high tertile of consumption. Thus, there are $3 \times 6 = 18$ possible exposure categories. The six genotypes, and number of subjects with that genotype in parentheses, are *1/*1 (180), *1/*3 (11), *1/*4 (80), *1/*5 (9), *4/*4 (10) and *1/*2 × 2 (12). Using the retrospective likelihood, Seaman & Richardson (2001) obtained the posterior mean log odds ratios and posterior standard deviations that are reported in Table 1 under the heading of ‘unconstrained retrospective’ model. An improper $\mathrm{Dir}(0, \ldots, 0)$ prior on the 18 exposure categories was assumed, but any Dirichlet prior could have been used. The baseline category is genotype *1/*1 and low tobacco consumption. It can be seen that, except for *1/*2 × 2, all genotypes appear more risky than the baseline genotype, and medium and high tobacco consumption appear to increase the risk by the same amount, compared to low consumption.

Table 1. *Results from lung cancer study. Posterior means and standard deviations*, sd, *for three models are presented. See text for description of the unconstrained, constrained and unconstrained continuous models*

	Exposure variable						
	CYP2D6 genotype					Smoking factor	
	*1/*3	*1/*4	*1/*5	*4/*4	*1/*2 × 2	medium	high
	Unconstrained retrospective model						
Posterior mean	1·04	0·63	0·30	0·89	−0·28	0·89	0·89
Posterior sd	0·70	0·28	0·74	0·72	0·68	0·30	0·30
	Constrained prospective model						
Posterior mean	1·04	0·58	0·27	1·28	−0·25	0·78	1·02
Posterior sd	0·69	0·26	0·74	0·55	0·68	0·28	0·28
	Unconstrained continuous model						
Posterior mean	0·87	0·64	0·18	0·86	−0·24	0·73[†]	
Posterior sd	0·69	0·28	0·74	0·72	0·68	0·21[†]	

[†] This is for the continous variable log tobacco consumption.

In order to ensure that Monte Carlo standard errors associated with all log odds ratio estimates were less than 0·005, it was necessary to generate a chain of length 1 000 000 iterations.

We fitted the equivalent prospective model, the ‘unconstrained prospective’ model, using WinBUGS vl.3. Naturally, the same estimates of log odds ratio were obtained. Mixing of the Markov chain produced was far superior to that of the retrospective model, because of the absence from the prospective model of parameters describing the exposure distribution. To ensure that all Monte Carlo standard errors were less than 0·005, a chain length of only 20 000 iterations was needed.

To indicate the potential of Bayesian modelling of case-control studies, we now describe a simple model that takes into account the structure of the two covariates in the dataset. It might be assumed that the risk associated with the heterozygous genotype *1/*4 would be intermediate to those of the two corresponding homozygous genotypes *1/*1 and *4/*4. Also, we may assume that the risk cannot decrease as tobacco consumption increases. It is simple to fit a model that incorporates these constraints. The estimates from this 'constrained prospective' model are also given in Table 1. As expected, compared to the estimates from the unconstrained model, the log odds ratios of *1/*4 and of medium tobacco consumption have decreased and those of *4/*4 and high tobacco consumption have increased. The log odds ratio estimate of genotype *4/*4 is affected more than that of *1/*4 because the latter genotype is more common and so its log odds ratio is estimated with more precision. The posterior standard deviations of all four parameters have decreased, showing how the use of the two assumptions has enabled risk parameters to be estimated more precisely.

This 'constrained prospective' model is a simple case of a more elaborate model we have developed for efficiently analysing case-control studies of highly polymorphic candidate genes (Seaman et al., 2002).

Finally, we fitted an unconstrained model that treats log tobacco consumption, in units of gram-years, as a continuous variable, rather than grouping it into tertiles. Results of this 'unconstrained continuous' model are shown in Table 1.

7. Discussion

The $\text{Dir}(a, \ldots, a)$ distribution with a tending to zero is an obvious choice for a noninformative prior on exposure, and is particularly suitable when the study includes continuous exposure variables. It is also attractive for two other reasons. First, as has been shown by Seaman & Richardson (2001), when the method described by Müller & Roeder (1997) for continuous exposure variables is adapted for use with categorical variables, the appropriate prior for exposure is the $\text{Dir}(0, \ldots, 0)$ density. Secondly, there is symmetry in the prior assumptions about exposure in cases and controls. If the prior on exposure in controls is $\text{Dir}(0, \ldots, 0)$ and independent of the prior on the log odds ratios, it implies that the prior on exposure in cases is also $\text{Dir}(0, \ldots, 0)$ and independent of the prior on the log odds ratios (Seaman & Richardson, 2001). However, there is a sense in which the $\text{Dir}(0, \ldots, 0)$ prior is not noninformative. Let δ_ε denote the measure which is zero everywhere except at $(0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in position ε of the vector. Sethuraman & Tiwari (1982) showed that, as a tends to zero, the $\text{Dir}(a, \ldots, a)$ density converges to a random degenerate measure δ_ε, where ε is drawn from the discrete uniform distribution on $\{1, \ldots, J\}$. Loosely speaking therefore, if $(\phi_1, \ldots, \phi_J) \sim \text{Dir}(a, \ldots, a)$, then, as a becomes increasing small, it is increasingly probable that one of $\phi_1, \ldots, \phi_J$ is very close to one and the rest are very close to zero. It is for this very reason that the 'limiting' prospective model does not depend on the assumed support, $\{z_1, \ldots, z_J\}$, of the exposure vector: as a tends to zero, any exposure values that are assumed possible but that are not observed in the study have posterior probabilities that tend to zero.

The problem of finding a model that works with small stratum sizes remains unsolved. In the classical frequentist analysis, conditional logistic regression may be used. Although approaches that use the conditional likelihood within a Bayesian framework have been described by Spiegelhalter et al. (1995) and Diggle et al. (2000) for pair-matched, or 1:*m*-matched, studies, they have not yet been justified theoretically. As Diggle et al. (2000) have noted, a correct Bayesian analysis ought to be based on the full likelihood, as is used in unconditional logistic regression and the prospective model of this paper. In an unpublished report, K. M. Rice showed that the conditional likelihood may be obtained from the full prospective likelihood by assuming a particular ‘invariant’ prior for the nuisance parameter and then integrating it out. However, this result is not immediately applicable to the analysis of case-control studies unless an invariant prior can be found that makes the prospective model equivalent to a retrospective model.

Acknowledgement

The authors wish to thank Wally Gilks, David Spiegelhalter, Ken Rice, Vern Farewell, Sonia Petrone and especially David Clayton for useful discussion and two anonymous referees for very helpful comments. Shaun Seaman thanks INSERM for giving him a Poste Vert fellowship to carry out this work.

Appendix

Propriety of the posterior distribution of the Poisson model

Here we prove that the posterior density, $p(\alpha, \beta, \delta|y)$, of the Poisson model is proper. It suffices to show that $p(\omega, \delta|y)$ is proper. Let

$$I(\delta)=\int_{-\infty}^{\infty}\prod_{j=1}^{J}\left\{\frac{1}{1+\exp(\omega+\delta^{\mathrm{T}}z_j)}\right\}^{y_{0j}+a_j}\left\{\frac{\exp(\omega+\delta^{\mathrm{T}}z_j)}{1+\exp(\omega+\delta^{\mathrm{T}}z_j)}\right\}^{y_{1j}}d\omega.$$

Choose q and r to be any integers in $\{1, \ldots, J\}$ such that $y_{0q} \geqslant 1$ and $y_{0r} \geqslant 1$. Then

$$\begin{aligned}I(\delta)&<\int_{-\infty}^{\infty}\left\{\frac{1}{1+\exp(\omega+\delta^{\mathrm{T}}z_q)}\right\}^{y_{0q}+a_q}\left\{\frac{\exp(\omega+\delta^{\mathrm{T}}z_r)}{1+\exp(\omega+\delta^{\mathrm{T}}z_r)}\right\}^{y_{1r}}d\omega\\
&<\int_{\delta^{\mathrm{T}}z_q}^{\infty}\frac{1}{1+\exp(\omega)}d\omega+\int_{-\infty}^{\delta^{\mathrm{T}}z_r}\frac{\exp(\omega)}{1+\exp(\omega)}d\omega\\
&=\int_{-\infty}^{-\delta^{\mathrm{T}}z_q}\frac{\exp(\omega)}{1+\exp(\omega)}d\omega+\int_{-\infty}^{\delta^{\mathrm{T}}z_r}\frac{\exp(\omega)}{1+\exp(\omega)}d\omega\\
&=\log\{1+\exp(-\delta^{\mathrm{T}}z_q)\}+\log\{1+\exp(\delta^{\mathrm{T}}z_r)\}.\end{aligned}$$

Thus,

$$\int I(\delta)p(\delta)d\delta<\int\log\{1+\exp(-\delta^{\mathrm{T}}z_q)\}p(\delta)d\delta+\int\log\{1+\exp(\delta^{\mathrm{T}}z_r)\}p(\delta)d\delta.$$

However,

$$\begin{aligned}\int\log\{1+\exp(-\delta^{\mathrm{T}}z_q)\}p(\delta)d\delta&<\int_{\{\delta:\,z_q^{\mathrm{T}}\delta>0\}}(\log 2)p(\delta)d\delta+\int_{\{\delta:\,z_q^{\mathrm{T}}\delta\leqslant 0\}}\log\{2\exp(-\delta^{\mathrm{T}}z_q)\}p(\delta)d\delta\\
&\leqslant\log 2+E(-z_q^{\mathrm{T}}\delta|z_q^{\mathrm{T}}\delta\leqslant 0).\end{aligned}$$

A similar result holds for r, and hence

$$\int I(\delta)p(\delta)d\delta < 2\log 2 + E(-z_q^{\mathrm{T}}\delta | z_q^{\mathrm{T}}\delta \leqslant 0) + E(z_r^{\mathrm{T}}\delta | z_r^{\mathrm{T}}\delta \geqslant 0).$$

Thus, if $-\infty < E(z_q^{\mathrm{T}}\delta) < \infty$ and $-\infty < E(z_r^{\mathrm{T}}\delta) < \infty$, then $\int I(\delta)p(\delta)d\delta < \infty$, and so $p(\omega, \delta|y)$ corresponds to a proper distribution.

References

Ashby, D., Hutton, J. L. & McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *Statistician* **42**, 385–97.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Statistician* **43**, 495–504.

Billingsley, P. (1986). *Probability and Measure*, 2nd ed. New York: Wiley.

Consonni, G. & Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *J. Am. Statist. Assoc.* **90**, 935–44.

Devroye, L. (1986). *Non-uniform Random Variate Generation.* New York: Springer-Verlag.

Diggle, P. J., Morris, S. E. & Wakefield, J. C. (2000). Point-source modelling using matched case-control data. *Biostatistics* **1**, 89–105.

George, E. I. & McCulloch, R. E. (1996). Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice*, Ed. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, pp. 203–14. London: Chapman and Hall.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Gustafson, P., Le, N. D. & Vallee, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3**, 229–43.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statist. Med.* **7**, 1223–30.

Müller, P. & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–37.

Müller. P., Parmigiani, G., Schildkraut, J & Tardella, L. (1999). Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55**, 858–66.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

Nurminen, M. & Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scand. J. Statist.* **14**, 67–77.

Pike, M. C., Hill, A. P. & Smith, P. G. (1980). Bias and efficiency in logistic analyses of stratified case-control studies. *Int. J. Epidem.* **9**, 89–95.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence model and case-control studies. *Biometrika* **66**, 403–11.

Raghunathan, T. E. (1994). Monte Carlo methods for exploring sensitivity to distributional assumptions in a Bayesian analysis of a series of 2 × 2 tables. *Statist. Med.* **13**, 1525–38.

Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). Semiparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Assoc.* **91**, 722–32.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130–4.

Seaman, S. R. & Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88**, 1073–88.

Seaman, S. R., Richardson, S., Stucker, I. & Benhamou, S. (2002). A Bayesian partition model for case-control studies on highly polymorphic candidate genes. *Genet. Epidem.* **22**, 356–368.

Sethuraman, J. & Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics III*, Ed. S. S. Gupta and J. O. Berger, pp. 305–15. London: Academic Press.

Spiegelhalter, D. J., Thomas. A., Best, N. G. & Gilks, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50.* Cambridge: Medical Research Council Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A. & Best, N. G. (1999). *WinBUGS Version 1.2 User Manual.* Cambridge: MRC Biostatistics Unit.

Zelen, M. & Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statist. Med.* **5**, 261–9.

[*Received April* 2002. *Revised April* 2003]