Biometrika (2015), **102**, 1, *pp.* 169–180
Printed in Great Britain

doi: 10.1093/biomet/asu048
Advance Access publication 2 December 2014

Using covariate-specific disease prevalence information to increase the power of case-control studies

By JING QIN

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

jingqin@niaid.nih.gov

HAN ZHANG

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.

han.zhang2@nih.gov

PENGFEI LI

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

pengfei.li@uwaterloo.ca

DEMETRIUS ALBANES and KAI YU

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.

albanesd@mail.nih.gov yuka@mail.nih.gov

Summary

Public registration databases and large cohort studies provide vital information on disease prevalence at various levels of a risk factor. This auxiliary information can be helpful in conducting statistical inference in a new study. We aim to develop a statistical procedure that improves the efficiency of the logistic regression model for a case-control study by utilizing auxiliary information on covariate-specific disease prevalence via a series of unbiased estimating equations. We adopt empirical likelihood for statistical inference, and demonstrate its advantages through simulation and an application.

Some key words: Auxiliary information; Biased sampling; Case and control studies; Empirical likelihood; Estimating equations; Meta-analysis.

1. Introduction

The case-control design provides a cost-efficient way of collecting information conditional on subjects' outcome status. It is widely used in epidemiological studies that identify risk factors for disease outcomes. Under such a retrospective design, information on risk factors and other covariates is collected for fixed numbers of cases and controls. Prospective likelihood based on a logistic regression model can be used to estimate the odds ratio owing to the equivalence of the prospective and retrospective maximum likelihood estimators (Prentice & Pyke, 1979). In many situations, in addition to data collected from the case-control study, we have information from previous studies or public databases. For example, in a case-control study of lung cancer, we might be interested in evaluating the interaction between smoking and certain genetic markers. In addition to the smoking information collected in the study, we might have access to the estimated disease risk for subjects with varying levels of smoking intensity, based on a larger study conducted in the same population. We aim to develop a unified framework to improve inference by integrating auxiliary information from external sources with that from the case-control study.

The use of auxiliary information is a popular strategy for increasing efficiency in statistical inference. For example, Imbens & Lancaster (1994) show that census reports provide nearly exact estimates of the moments of the marginal distribution of economic variables, and can be used in combination with cross-sectional or panel samples to improve estimation accuracy in economic studies. The integration of auxiliary information into survey sampling has been studied extensively (e.g., Cochran, 1997; Särndal et al., 1992). Chatterjee & Carroll (2005) use another type of auxiliary information, i.e., the fact that the gene and environment are independent in the general population, to improve power for the detection of a gene-environment interaction.

In this paper we aim to integrate auxiliary information into the analysis of case-control data. We assume that we know disease prevalence at various levels of a known risk factor X from a database, a published meta-analysis or a large cohort study. We are interested in studying the effect of a candidate risk factor Y, or the interaction between X and Y, incorporating auxiliary information on X, under the logistic regression model. Zaitlen et al. (2012) proposed such a procedure for population-based genetic association studies using the liability threshold model as the disease risk model. This model is common in population genetics, but is rarely used in epidemiological studies, in part because of its lack of interpretability. We use the logistic regression model as the disease risk model. Auxiliary information, such as the disease prevalence at different levels of X, can be expressed as constraints on the regression coefficients and the joint covariate distribution, so we use empirical likelihood for general estimating equations (Qin & Lawless, 1994) to improve the efficiency of the logistic regression model. With some modifications, this procedure can also be used to analyse a case-control study nested within a large prospective cohort study from which the disease prevalence is derived at various levels of the risk factor X. In this scenario, the auxiliary information is not independent of the case-control data.

2. Main results

Let $D=1$ or 0 represent diseased or disease-free status of a subject, and let X and Y be the two risk factors. The typical model for a binary response given X and Y is the logistic regression model,

$$\mathrm{pr}(D=1 \mid x, y)=\frac{\exp(\alpha+x\beta+\gamma y+\xi xy)}{1+\exp(\alpha+x\beta+\gamma y+\xi xy)},$$

where β and γ are log odds ratio parameters for X and Y, respectively, and ξ characterizes the interaction between X and Y.

Instead of prospectively collecting (D, X, Y), in case-control studies typically one collects (X, Y) by conditioning on the status of D. This is retrospective or case-control sampling. Let (X_{i1}, Y_{i1}) $(i=1,\ldots,n_1)$ be the covariate data for the cases with $D_i=1$, where n_1 is a prespecified number, and, let (X_{i0}, Y_{i0}) $(i=1,\ldots,n_0)$ be the covariate data for the controls with $D_i = 0$. Again, n_0 is a fixed number. Using Bayes' formula, we have

$$f(x, y \mid D = 1) = \frac{\text{pr}(D = 1 \mid x, y) f(x, y)}{\text{pr}(D = 1)}, \quad f(x, y \mid D = 0) = \frac{\text{pr}(D = 0 \mid x, y) f(x, y)}{\text{pr}(D = 0)},$$

where $f(x, y)$ is the joint density of (X, Y). The case and control densities are linked by the exponential tilting model

$$f(x, y \mid D = 1) = \exp(\alpha^* + x\beta + \gamma y + \xi xy) f(x, y \mid D = 0), \tag{1}$$

where $\alpha^* = \log\{(1 - \pi)/\pi\} + \alpha$, with $\pi = \text{pr}(D = 1)$ the overall disease prevalence (Qin & Zhang, 1997).

Without any auxiliary information on $f(x, y \mid D = 0)$, Prentice & Pyke (1979) showed that one may perform a prospective logistic likelihood analysis with loglikelihood

$$\ell_P = \sum_{i=1}^{n_1} (\alpha^* + \beta X_{i1} + \gamma Y_{i1} + \xi X_{i1} Y_{i1}) - \sum_{i=1}^{n} \log\{1 + \exp(\alpha^* + \beta X_i + \gamma Y_i + \xi X_i Y_i)\},$$

where we have used (X_i, Y_i) $(i = 1, \ldots, n)$, where $n = n_0 + n_1$, to denote the pooled data, and we have assumed that $D_i = 1$ $(i = 1, \ldots, n_1)$ and $D_i = 0$ $(i = n_1 + 1, \ldots, n)$.

We assume that X is a covariate whose effect on the outcome has been well-studied, such as smoking or body mass index. Let Y be the new covariate under study, such as a genotype measured on a genetic marker. Based on the published information we assume that the disease prevalence at various levels of X, i.e.,

$$\text{pr}(D = 1 \mid a < X \leqslant b) = \phi(a, b), \tag{2}$$

is known. Next we demonstrate that this published information can be transformed into an unbiased estimating equation, which will improve the estimation of the unknown parameters.

Recall that $\pi = \text{pr}(D = 1)$ and let $F_1(x)$ and $F_0(x)$ respectively denote the cumulative distribution functions of X for the case and control populations. Using Bayes' formula, we have

$$\frac{\pi \int_a^b \mathrm{d}F_1(x)}{\text{pr}(a < X \leqslant b)} = \phi(a, b), \quad \frac{(1 - \pi) \int_a^b \mathrm{d}F_0(x)}{\text{pr}(a < X \leqslant b)} = 1 - \phi(a, b),$$

giving

$$\int_a^b \mathrm{d}F_1(x) = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} \int_a^b \mathrm{d}F_0(x),$$

or equivalently

$$E_1\{I(a < X \leqslant b)\} = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} E_0\{I(a < X \leqslant b)\},$$

where E_1 and E_0 denote expectations with respect to the case and control populations, respectively. Using (1), we have

$$E_0\{I(a < X \leqslant b) \exp(\alpha^* + \beta X + \gamma Y + \xi XY)\} = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} E_0\{I(a < X \leqslant b)\},$$

or equivalently

$$E_0\left[I(a < X \leqslant b)\left\{\exp(\alpha^* - \eta + \beta X + \gamma Y + \xi XY) - \frac{\phi(a,b)}{1-\phi(a,b)}\right\}\right] = 0, \tag{3}$$

where $\eta = \log\{(1-\pi)/\pi\}$. Thus, the published information in (2) is transformed into the unbiased estimating equation in (3).

Let $a_0 < a_1 < \cdots < a_I$ be a partition of the X space, and let $\theta = (\eta, \alpha^*, \beta, \gamma, \xi)^{\mathrm{T}}$ be a vector containing all the unknown parameters. Denote

$$g_i(X, Y; \theta) = I(a_{i-1} < X \leqslant a_i)\left\{\exp(\alpha^* - \eta + \beta X + \gamma Y + \xi XY) - \frac{\phi(a_{i-1}, a_i)}{1-\phi(a_{i-1}, a_i)}\right\}.$$

Note that $\phi(a_{i-1}, a_i)$ $(i = 1, \ldots, I)$ is the auxiliary information extracted from the published data. Further, let $g(X, Y; \theta) = \{g_1(X, Y; \theta), \ldots, g_I(X, Y; \theta)\}^{\mathrm{T}}$. By (3), the published information can be summarized in the unbiased estimating equations

$$E_0\{g(X, Y; \theta)\} = 0. \tag{4}$$

We now use empirical likelihood to incorporate the auxiliary information in the estimating equations (4) into our estimation of the unknown parameters. Let

$$\delta(X, Y; \theta) = \exp(\alpha^* + \beta X + \gamma Y + \xi XY).$$

Then $E_0\{\delta(X, Y; \theta) - 1\} = 0$ by (1). Let the joint cumulative distribution functions of (X, Y) for the case and control groups be $F_1(x, y)$ and $F_0(x, y)$. By discretizing $F_0(x, y)$ at each of the observed data points, we obtain the log empirical likelihood of θ and $F_0(x, y)$,

$$\tilde{\ell} = \sum_{i=1}^{n} D_i(\alpha^* + \beta X_i + \gamma Y_i + \xi X_i Y_i) + \sum_{i=1}^{n} \log p_i,$$

with $p_i = dF_0(X_i, Y_i)$ satisfying the constraints

$$p_i \geqslant 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i\{\delta(X_i, Y_i; \theta) - 1\} = 0, \quad \sum_{i=1}^{n} p_i g(X_i, Y_i; \theta) = 0. \tag{5}$$

The first two constraints in (5) ensure that $F_0(x, y)$ is a proper cumulative distribution function. The third constraint corrects the biased sampling in the case-control study and guarantees that $F_1(x, y)$ is also a proper cumulative distribution function. The last constraint uses the auxiliary information in (4) to improve the efficiency. In the setting of random sampling, Qin & Lawless (1994) used Owen (1988)'s empirical likelihood method to incorporate auxiliary information. The current set-up generalizes their method to biased sampling or case-control studies.

If the overall disease prevalence $\pi = \mathrm{pr}(D = 1)$ is known, then η is known. If π is unknown, then, provided $I = 1$, π is still identifiable. If $I > 1$, then we have more estimating equations than unknown parameters. In the following we will consider the scenario where η is unknown and $I \geqslant 1$. The situation where η is known can be dealt with by substituting the known value into the estimating equations.

By the Lagrange multiplier method, the profile likelihood of θ is

$$\ell(\theta)=\sum_{i=1}^{n} D_i(\eta+\alpha+\beta X_i+\gamma Y_i+\xi X_i Y_i) \\ -\sum_{i=1}^{n}\log[1+\lambda\{\delta(X_i,Y_i;\theta)-1\}+t^{\mathrm{T}}g(X_i,Y_i;\theta)],$$

where the Lagrange multipliers λ and $t=(t_1,\ldots,t_I)^{\mathrm{T}}$ are determined by

$$\sum_{i=1}^{n}\frac{\delta(X_i,Y_i;\theta)-1}{1+\lambda\{\delta(X_i,Y_i;\theta)-1\}+t^{\mathrm{T}}g(X_i,Y_i;\theta)}=0,$$

$$\sum_{i=1}^{n}\frac{g(X_i,Y_i;\theta)}{1+\lambda\{\delta(X_i,Y_i;\theta)-1\}+t^{\mathrm{T}}g(X_i,Y_i;\theta)}=0.$$

The true values of λ and t are $\lambda_0=n_1/n$ and 0, respectively; see Qin & Zhang (1997) and Qin & Lawless (1994). For point estimation, we can maximize $\ell(\theta)$ with respect to θ to obtain the maximum empirical likelihood estimator of θ, i.e., $\hat\theta=\arg\sup_\theta \ell(\theta)$. Let $\hat\lambda$ and $\hat t$ be the Lagrange multipliers corresponding to $\hat\theta$. Then $p_i=\mathrm{d}F_0(X_i,Y_i)$ and $q_i=\mathrm{d}F_1(X_i,Y_i)$ can be estimated by

$$\hat p_i=\frac{1}{n}\frac{1}{1+\hat\lambda\{\delta(X_i,Y_i;\hat\theta)-1\}+\hat t^{\mathrm{T}}g(X_i,Y_i;\hat\theta)},\quad \hat q_i=\hat p_i\delta(X_i,Y_i;\hat\theta)\quad (i=1,\ldots,n). \tag{6}$$

Suppose that the true value of θ is $\theta_0=(\eta_0,\alpha_0^*,\beta_0,\gamma_0,\xi_0)^{\mathrm{T}}$. In the Appendix and Supplementary Material, we prove the following theorems.

Theorem 1. *Suppose that $\rho=n_1/n_0$ remains constant as $n\to\infty$ and $\rho\in(0,1)$. Under regularity conditions, as n goes to infinity, $n^{1/2}(\hat\theta-\theta_0)\to N(0,\Sigma)$ in distribution, where $\Sigma=J^{-1}-\rho^{-1}(1+\rho)^2(1,1,0,0,0)^{\mathrm{T}}(1,1,0,0,0)$ and $J=UV^{-1}U^{\mathrm{T}}$. The matrices U and V are defined in* (A4) *of the Appendix.*

Similarly to Qin & Lawless (1994) the estimation of the logistic regression coefficients $\theta_2=(\beta,\gamma,\xi)^{\mathrm{T}}$ becomes more precise as the number of estimating equations increases.

Corollary 1. *Let $\theta_{02}=(\beta_0,\gamma_0,\xi_0)^{\mathrm{T}}$ be the true value of θ_2 and let $\hat\theta_{2L}$ be the maximum likelihood estimator of θ_2 based on logistic regression in the absence of auxiliary information. Under the conditions of Theorem* 1, *we have:*

(a) *if $I=1$, the asymptotic variance of $n^{1/2}(\hat\theta_2-\theta_{02})$ is the same as that of $n^{1/2}(\hat\theta_{2L}-\theta_{02})$;*
(b) *if $I>1$, the difference of the asymptotic covariance matrices of $n^{1/2}(\hat\theta_2-\theta_{02})$ and $n^{1/2}(\hat\theta_{2L}-\theta_{02})$ is nonpositive-definite; and*
(c) *when $I>1$, the asymptotic covariance matrix of $n^{1/2}(\hat\theta_2-\theta_{02})$ cannot decrease if an estimating equation in* (4) *is dropped.*

When testing $\xi=0$, the semiparametric likelihood ratio statistic is $R(0)=2\left\{\sup_\theta \ell(\theta)-\sup_{\eta,\alpha^*,\beta,\gamma,\xi=0}\ell(\theta)\right\}$. We can show the result below.

Theorem 2. *Under the conditions of Theorem* 1, *as n goes to infinity, the empirical likelihood ratio statistic $R(0)\to\chi_1^2$ in distribution if $\xi=0$.*

Remark 1. The auxiliary information summarized in (3) should be informative for estimating β and ξ, but not for estimating γ. This can be observed through the following equation:

$$\int I(a < x \leqslant b)\delta(x, y; \theta)\, \mathrm{dF}_0(y, x) = \int I(a < x \leqslant b) \exp(\alpha^* + \beta x + s + \xi xs/\gamma)\, \mathrm{dF}_0(s/\gamma, x).$$

Since the underlying distribution $F_0(x, y)$ is not specified, we can treat $F_0(s/\gamma, x)$ as a new underlying distribution $F_0^*(s, x)$. After profiling out $F_0^*(s, x)$, we observe that the auxiliary information equation does not involve γ if $\xi = 0$. Even if $\xi \neq 0$, the information for γ is minimal since γ and ξ are entangled.

Remark 2. We have assumed that the auxiliary information $\phi(a, b)$ in the estimating equation (3) is either precise or comes from a large separate study with sample size N, where $n/N \to 0$. However, we must consider the variation in the auxiliary information $\phi(a, b)$ when the outside information comes from an independent source with a limited sample size, Case I. In general it can be shown that the likelihood ratio statistic $R(0)$ has an asymptotic scaled chi-squared distribution. When the outside information comes from the cohort from which the case-control data are drawn, called Case II, then we must take into account the correlation between the auxiliary information and the case-control data. Again, it can be shown that the likelihood ratio statistic $R(0)$ has an asymptotic scaled chi-squared distribution. We can calculate the scale parameter using the asymptotic formula, but the following bootstrap procedure provides an accurate estimate for the scale parameters in Case II for testing the interaction. A similar bootstrap method can be applied with mild adjustment for Case I. In the procedure below, we use $D_1, \ldots, D_N$ to denote the disease outcomes of subjects from the full cohort.

Step 1. Based on the original case-control data and auxiliary information $\phi(a_i, b_i)$ $(i = 1, \ldots, I)$, calculate the test statistic $R(0)$. Let $\tilde{\theta}$ be the maximum empirical likelihood estimate of θ under the null hypothesis $H_0 : \xi = 0$. The corresponding Lagrange multipliers are denoted by $\tilde{\lambda}$ and $\tilde{t}$. Similarly to (6), under H_0, we estimate $p_i = \mathrm{d}F_0(X_i, Y_i)$ and $q_i = \mathrm{d}F_1(X_i, Y_i)$ by

$$\tilde{p}_i = \frac{1}{n} \frac{1}{1 + \tilde{\lambda}\{\delta(X_i, Y_i; \tilde{\theta}) - 1\} + \tilde{t}^{\mathrm{T}} g(X_i, Y_i; \tilde{\theta})}, \quad \tilde{q}_i = \tilde{p}_i \delta(X_i, Y_i; \tilde{\theta}) \quad (i = 1, \ldots, n).$$

We further estimate $F_0(x, y)$ and $F_1(x, y)$ by

$$\tilde{F}_0(x, y) = \sum_{i=1}^{n} \tilde{p}_i I(X_i \leqslant x; Y_i \leqslant y), \quad \tilde{F}_1(x, y) = \sum_{i=1}^{n} \tilde{q}_i I(X_i \leqslant x; Y_i \leqslant y).$$

Step 2. Resample $D_1, \ldots, D_N$ with replacement N times, denoted by $D_1^*, \ldots, D_N^*$. If $D_i^* = 1$ then sample (X_{i1}^*, Y_{i1}^*) from $\tilde{F}_1(x, y)$ for this diseased subject; otherwise, sample (X_{i0}^*, Y_{i0}^*) using $\tilde{F}_0(x, y)$ for this disease-free subject.

Step 3. From the regenerated cohort data, randomly choose n_1 case data from subjects with $D_i^* = 1$, denoted by $(X_{i1}^{**}, Y_{i1}^{**})$ $(i = 1, \ldots, n_1)$. Similarly, from the resampled control data $D_i^* = 0$ in Step 2, randomly choose n_0 control data, denoted by $(X_{i0}^{**}, Y_{i0}^{**})$ $(i = 1, \ldots, n_0)$.

Step 4. Based on the regenerated cohort data in Step 2, calculate $\phi^*(a_i, b_i)$ $(i = 1, \ldots, I)$.

Step 5. Based on the case and control data selected in Step 3 and $\phi^*(a_i, b_i)$ $(i = 1, \ldots, I)$ calculated in Step 4, find the likelihood ratio statistic $R^b(0)$.

Table 1. *Type* I *error* (%) *and power* (%) *comparison for four scenarios: the results for each scenario are based on* 1000 *simulated datasets, each consisting of* 2000 *cases and* 2000 *controls with* $(\beta, \gamma) = (1{\cdot}00, 0{\cdot}08)$

	Type I error (%): $\xi = 0$				Power (%): $\xi = 0{\cdot}05$			
Scenario	1	2	3	4	1	2	3	4
Logistic	5·0	4·4	5·0	4·9	29·7	30·5	29·0	30·6
Proposed	5·0	5·2	5·6	5·2	91·0	94·4	50·3	51·9

Step 6. Repeat steps 2–5 B times. Use the bootstrap likelihood ratio statistics $R^b(0)$ $(b = 1, \ldots, B)$ to calibrate the scale parameter as $B^{-1}\sum_{b=1}^{B} R^b(0)$. Then rescale the original likelihood ratio statistic $R(0)$ by this scaling parameter and refer it to the standard chi-squared distribution for the p-value.

Our simulation results indicate that this bootstrap method can control the Type I error properly.

3. Simulation studies

In each of the four simulation scenarios considered here, we tested the null hypothesis $\xi = 0$, at the nominal level 0·05. We compared our method with the likelihood ratio test based on standard logistic regression without any auxiliary information. All the simulations were conducted with 2000 cases and 2000 controls and averaged over 1000 replications. We first considered the cases where the auxiliary information is assumed known without any uncertainty, and thus the proposed test statistic follows the χ^2_1 distribution. We then considered the cases where the auxiliary information was estimated from a cohort with a limited sample size. In these cases, the test statistic follows a scaled χ^2_1 distribution. We estimated the scale parameter using the proposed bootstrap procedure with $B = 500$. We report the Type I errors and powers in Table 1, and we compare the estimated bias and standard deviation in Table 2.

First, we considered the situation where only the prevalences of the given levels of X are available, called Scenario 1. In the control group, we generate X and Y as follows:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu_0, \Sigma_0), \quad \mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We set $\rho = 0{\cdot}5$ to allow moderate correlation between the two variables. We generated n_0 controls from $N(\mu_0, \Sigma_0)$. Using the exponential tilting model (1), we generated n_1 cases from another normal distribution with the density $\exp(\alpha^* + \beta x + \gamma y + \xi xy)\phi(x, y; \mu_0, \Sigma_0)$, where $\phi(\cdot)$ is the density function of the bivariate normal distribution $N(\mu_0, \Sigma_0)$. The auxiliary information on the disease prevalence was assumed to be known for X in the intervals $(-\infty, -0{\cdot}67]$, $(-0{\cdot}67, 0]$, $(0, 0{\cdot}67]$, and $(0{\cdot}67, \infty)$, where $\pm 0{\cdot}67$ are the first and third quartiles of the standard normal distribution. Table 2 shows that the use of the auxiliary information can greatly improve the estimation of the coefficient of X, i.e., β. In our setting, the variance ratio between the new estimator and that from the standard logistic regression model can be as low as 25%. We also observe a noticeable variance reduction in the estimated interaction coefficient ξ. The variance reduction in the estimate of γ is negligible, as explained in Remark 1.

We then considered the situation where auxiliary information for the second variable Y was also available, called Scenario 2, i.e.,

$$\mathrm{pr}(D = 1 \mid b_{k-1} < Y \leqslant b_k) = \psi(b_{k-1}, b_k) \quad (k = 1, \ldots, K),$$

Table 2. *Bias and standard deviation (in parentheses) comparison for four scenarios: the results for each scenario are* 1000× *the actual values and are based on* 1000 *simulated datasets, each consisting of* 2000 *cases and* 2000 *controls with* $(\beta, \gamma, \xi) = (1{\cdot}00, 0{\cdot}08, 0{\cdot}05)$

Scenario	$\beta = 1{\cdot}00$		$\gamma = 0{\cdot}08$		$\xi = 0{\cdot}05$	
	Logistic	Proposed	Logistic	Proposed	Logistic	Proposed
1	2 (46)	0 (23)	−1 (46)	0 (43)	1(37)	−3 (14)
2	2 (45)	0 (11)	2 (44)	0 (20)	1 (36)	1 (15)
3	1 (44)	−4 (27)	1 (44)	9 (50)	1 (36)	−1 (26)
4	2 (45)	−4 (28)	2 (44)	9 (50)	1 (36)	1 (25)

were known. The cut-off points for X were $-0{\cdot}67$, 0, and $0{\cdot}67$, and those for Y were $\pm 0{\cdot}431$. There are large improvements in the estimates of all the coefficients (β, γ, ξ).

Finally, we conducted simulation studies in situations where the auxiliary information was estimated with some uncertainty. We considered two scenarios. Scenario 3 considered a case-control study with 2000 cases and 2000 controls, sampled from a cohort study with a sample size of 100 000, with auxiliary information estimated from the same cohort study. In Scenario 4 the auxiliary information was estimated from a separate cohort study, independent of the case-control study. The other parameters were similar to those used in Scenario 1. Table 1 shows that the bootstrap procedure can control the Type I error appropriately in both scenarios. We emphasize that this adjustment is critical to ensure the proper control of the Type I error. For example, the Type I errors for Scenarios 3 and 4 increase to $0{\cdot}290$ and $0{\cdot}273$, respectively, if the χ^2_1 distribution is used to evaluate the p-values. After the proper adjustment of the scale parameter, the proposed procedure had a clear power advantage over the standard logistic regression model, although the improvement was not as large as that observed in Scenario 1, where the auxiliary information was assumed known.

We studied the sensitivity of our test in situations where the auxiliary information was given incorrectly; see the Supplementary Material. We also conducted simulation studies in the setting of genetic association, where the variable Y represents the genotype, and X represents an environmental risk factor. The conclusions are similar to those reported in Tables 1 and 2.

4. Real-data application

In this section we apply our method to gene-smoking interaction in a case-control study of lung cancer. Cigarette smoking is a major risk factor for lung cancer. Recent genome-wide association studies have identified a few chromosomal regions, e.g., chromosomes 15q25, 15q25, 5p15, and 6p21, harbouring genetic variants called SNPs, i.e., single nucleotide polymorphisms, underlying susceptibility to lung cancer (Amos et al., 2008; McKay et al., 2008; Thorgeirsson et al., 2008; Wang et al., 2008; Landi et al., 2009; Rafnar et al., 2009). In particular, the chromosome 15q25 region has been shown to be associated with both lung-cancer risk and smoking behaviour. It is of great interest to test whether there is any interaction between the genetic variants in 15q25 and smoking. We applied our method to assess the evidence for interaction in a case-control study conducted within the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study, which was a randomized prevention trial including 29 133 male smokers in Finland between 1985 and 1988 (ATBC Cancer Prevention Study Group, 1994; Albanes et al., 1996). The case-control study consisted of 1694 lung-cancer cases and 1261 controls. We focused on the average intensity of cigarette smoking in terms of the average number of cigarettes per day, denoted by X, and evaluated its interaction with each of the 36 relatively common SNPs within the 15q25 region.

Table 3. *p-values (%) for the interaction test $H_0 : \xi = 0$ in the lung-cancer data*

SNP	rs1062980	rs13180	rs12903150	rs3743079	rs8192475	rs12914385	rs2036534
Proposed	1·2	1·3	1·9	3·4	7·9	8·8	8·8
Logistic	2·8	2·9	1·6	10·0	9·7	28·0	20·0

All the genotypes were measured as part of a lung cancer genome-wide association study (Landi et al., 2009).

Besides the information on smoking and SNPs measured in the case-control study, we also obtained information on lung cancer prevalence at various levels of X from the Alpha-Tocopherol, Beta-Carotene Study. This external information can be summarized as follows: $\phi(0, 10) = 8{\cdot}3\%$, $\phi(10, 15) = 10{\cdot}9\%$, $\phi(15, 20) = 12{\cdot}8\%$, $\phi(20, 25) = 14{\cdot}9\%$, $\phi(25, \infty) = 16{\cdot}6\%$. There are 4792, 4847, 9839, 4272, and 5383 subjects in each of the 5 groups.

In the analysis, we adjusted for age, grouped in 10-year intervals, and the second principal component for the control of population stratification (Price et al., 2006). We applied the Box–Cox transformation to X to find how to model its main effect. The log transformation was chosen because it maximized the likelihood of the model that consisted of age, the second principal component, and X on a transformed scale. The interaction between a candidate SNP and X was modelled as a product term between the genotype, coded as 0, 1, and 2, and $\log(X)$. To account for the uncertainty in disease prevalence at various levels of X estimated from the Alpha-Tocopherol, Beta-Carotene Study, we used 500 bootstrap samples to estimate the scaling parameter for the scaled chi-squared distribution under the null.

Table 3 shows the results for seven SNPs with p-values less than 0·1 from at least one considered test. Among the 36 SNPs, our method found four SNPs with p-values less than 0·05, while the likelihood ratio test based on the standard logistic regression model detected three. The two tests generated similar results, but the results from our method appeared to be slightly more significant, except for one SNP, rs12903150. Since the study from which the auxiliary information was derived was not large, there was considerable variation in the prevalence estimates, which could limit the power of our method. The advantages of our method would be more obvious if the auxiliary information were measured more precisely.

Acknowledgement

We thank the editor, associate editor and referees for constructive comments that greatly improved the article.

Supplementary material

Supplementary material available at *Biometrika* online contains proofs of Theorem 1, Corollary 1, Theorem 2, and more simulations.

Appendix

Outline of proof of Theorem 1

Let

$$l(\theta, t, \lambda) = \sum_{i=1}^{n} D_i(\alpha^* + \beta X_i + \gamma Y_i + \xi X_i Y_i) - \sum_{i=1}^{n} \log\left[1 + \lambda\{\delta(X_i, Y_i; \theta) - 1\} + t^{\mathrm{T}} g(X_i, Y_i; \theta)\right].$$

Then $\ell(\theta)=l(\theta,t,\lambda)$ with t and λ being the solutions to $\partial l(\theta,t,\lambda)/\partial t=0$ and $\partial l(\theta,t,\lambda)/\partial\lambda=0$. Further, let $\omega=(\theta^{\mathrm{T}},t^{\mathrm{T}})^{\mathrm{T}}$ and $\hat\omega=(\hat\theta^{\mathrm{T}},\hat t^{\mathrm{T}})^{\mathrm{T}}$. We have that $\hat\omega=(\hat\theta^{\mathrm{T}},\hat t^{\mathrm{T}})^{\mathrm{T}}$ and $\hat\lambda$ should satisfy

$$\frac{\partial l(\hat\theta,\hat t,\hat\lambda)}{\partial\omega}=0,\qquad \frac{\partial l(\hat\theta,\hat t,\hat\lambda)}{\partial\lambda}=0,$$

which together imply that $\hat\lambda=\lambda_0=n_1/n$; see the Supplementary Material for the details.

Applying the first-order Taylor expansion on $\partial l(\hat\theta,\hat t,\lambda_0)/\partial\omega$ and using the law of large numbers, we get

$$0=\frac{\partial l(\hat\theta,\hat t,\lambda_0)}{\partial\omega}=\frac{\partial l(\theta_0,0,\lambda_0)}{\partial\omega}+E\left\{\frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\omega\partial\omega^{\mathrm{T}}}\right\}(\hat\omega-\omega_0)+o_p(n^{1/2}).$$

Here $\omega_0=(\theta_0^{\mathrm{T}},0)^{\mathrm{T}}$ denotes the true value of ω. Then

$$\hat\omega-\omega_0=-\left[E\left\{\frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\omega\partial\omega^{\mathrm{T}}}\right\}\right]^{-1}\frac{\partial l(\theta_0,0,\lambda_0)}{\partial\omega}+o_p(n^{-1/2}). \tag{A1}$$

After some calculus and algebra, we show in Lemma 2 of the Supplementary Material that

$$\begin{aligned}E\left\{\frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\omega\partial\omega^{\mathrm{T}}}\right\}&=E\begin{pmatrix}\frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta^2} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial\alpha^*} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial\theta_2^{\mathrm{T}}} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial t^{\mathrm{T}}}\\ \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial\alpha^*} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial(\alpha^*)^2} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\alpha^*\partial\theta_2^{\mathrm{T}}} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\alpha^*\partial t^{\mathrm{T}}}\\ \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial\theta_2} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\alpha^*\partial\theta_2} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\theta_2\partial\theta_2^{\mathrm{T}}} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\theta_2\partial t^{\mathrm{T}}}\\ \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\eta\partial t} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial\alpha^*\partial t} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial t\partial\theta_2^{\mathrm{T}}} & \frac{\partial^2 l(\theta_0,0,\lambda_0)}{\partial t\partial t^{\mathrm{T}}}\end{pmatrix}\\ &=n\begin{pmatrix}0&0&0&-A_{14}\\0&-A_{22}&-A_{23}&-A_{24}\\0&-A_{32}&-A_{33}&-A_{34}\\-A_{41}&-A_{42}&-A_{43}&A_{44}\end{pmatrix}.\end{aligned} \tag{A2}$$

Lemma 2 gives the detailed forms of the constant A-matrices.

From (A1) and (A2), we can obtain the approximation of $\hat\theta-\theta_0$ after some tedious algebra work:

$$\hat\theta-\theta_0=J^{-1}UV^{-1}\big(n^{-1}S_n\big)+o_p(n^{-1/2}) \tag{A3}$$

with

$$S_n=\begin{pmatrix}\frac{\partial l(\theta_0,0,\lambda_0)}{\partial\alpha^*}\\ \frac{\partial l(\theta_0,0,\lambda_0)}{\partial\theta_2}\\ \frac{\partial l(\theta_0,0,\lambda_0)}{\partial t}\end{pmatrix},\quad U=\begin{pmatrix}0&0&A_{14}\\A_{22}&A_{23}&A_{24}\\A_{32}&A_{33}&A_{34}\end{pmatrix},\quad V=\begin{pmatrix}A_{22}&A_{23}&0\\A_{32}&A_{33}&0\\0&0&A_{44}\end{pmatrix},\quad J=UV^{-1}U^{\mathrm{T}}. \tag{A4}$$

In Lemma 3 of the Supplementary Material, we find that

$$E(S_n)=0,\quad \mathrm{var}(n^{-1/2}S_n)=V-\begin{pmatrix}A_{22}\\A_{32}\\A_{42}+A_{41}\end{pmatrix}\begin{pmatrix}A_{22}\\A_{32}\\A_{42}+A_{41}\end{pmatrix}^{\mathrm{T}}.$$

Note that S_n is a sum of independent random variables. By the central limit theorem, we conclude that

$$n^{1/2}(\hat{\theta}-\theta_0)=J^{-1}UV^{-1}\left(n^{-1/2}S_n\right)+o_p(1)\rightarrow N(0,\,\Sigma)$$

in distribution with $\Sigma=\{J^{-1}UV^{-1}\}\{\mathrm{var}(n^{-1/2}S_n)\}\{V^{-1}U^{\mathrm{T}}J^{-1}\}$. After some algebra, we can show that $\Sigma=J^{-1}-\rho^{-1}(1+\rho)^2(1,1,0,0,0)^{\mathrm{T}}(1,1,0,0,0)$, as claimed in Theorem 1.

Outline of proof of Corollary 1

Part (a). When $I=1$, U is a square matrix. Then (A3) implies that

$$U^{\mathrm{T}}(\hat{\theta}-\theta_0)=n^{-1}S_n+o_p(n^{-1/2}),$$

which can be used to find the following approximation of $(\hat{\alpha}^*-\alpha_0^*,\,\hat{\theta}_2^{\mathrm{T}}-\theta_{02}^{\mathrm{T}})^{\mathrm{T}}$:

$$\begin{pmatrix}\hat{\alpha}^*-\alpha_0^*\\ \hat{\theta}_2-\theta_{02}\end{pmatrix}=n^{-1}\begin{pmatrix}A_{22} & A_{23}\\ A_{32} & A_{33}\end{pmatrix}^{-1}\begin{pmatrix}\dfrac{\partial l(\theta_0,0,\lambda_0)}{\partial\alpha^*}\\ \dfrac{\partial l(\theta_0,0,\lambda_0)}{\partial\theta_2}\end{pmatrix}+o_p(n^{-1/2}). \qquad \text{(A5)}$$

Let $(\hat{\alpha}_L^*,\,\hat{\theta}_{2L}^{\mathrm{T}})^{\mathrm{T}}$ be the maximum likelihood estimator of $(\alpha_L^*,\,\theta_{2L}^{\mathrm{T}})^{\mathrm{T}}$ based on the logistic regression model. Qin & Zhang (1997) showed that $(\hat{\alpha}_L^*-\alpha_0^*,\,\hat{\theta}_{2L}^{\mathrm{T}}-\theta_{02}^{\mathrm{T}})^{\mathrm{T}}$ also has the approximation in (A5). Hence, the asymptotic variances of $n^{1/2}(\hat{\theta}_2-\theta_{02})$ and $n^{1/2}(\hat{\theta}_{2L}-\theta_{02})$ are the same.

Parts (b, c). The idea of the proof is similar to that of Corollary 1 in Qin & Lawless (1994) and is therefore omitted. See the Supplementary Material for more details.

References

Albanes, D., Heinonen, O. P., Taylor, P. R., Virtamo, J., Edwards, B. K., Rautalahti, M., Hartman, A. M., Palmgren, J., Freedman, L. S., Haapakoski, J., Barrett, M. J., Pietinen, P., Malila, N., Tala, E., Liippo, K., Salomaa, E. R., Tangrea, J. A., Teppo, L., Askin, F. B., Taskinen, E., et al. (1996). Alpha-Tocopherol and beta-carotene supplements and lung cancer incidence in the alpha-tocopherol, beta-carotene cancer prevention study: effects of base-line characteristics and study compliance. *J. Nat. Cancer Inst.* **88**, 1560–70.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y. Y., Chen, W. V., Shete, S., Spitz, M. R. & Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40**, 616–22.

ATBC Cancer Prevention Study Group (1994). The alpha-tocopherol, beta-carotene lung cancer prevention study: Design, methods, participant characteristics, and compliance. *Ann. Epidemiol.* **4**, 1–10.

Chatterjee, N. & Carroll, R. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**, 399–418.

Cochran, W. G. (1977). *Sampling Techniques.* New York: Wiley.

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Rev. Econ. Stud.* **61**, 655–80.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679–91.

McKay, J. D., Hung, R. J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., McLaughlin, J., Shepherd, F., Montpetit, A., Narod, S., et al. (2008). Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–6.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–49.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–12.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–9.

Qin, J. & Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

Qin, J. & Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609–18.

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S., Aben, K. K., Blöndal, T., Thorgeirsson, T. E., Thorleifsson, G., Kristjansson, K., Thorisdottir, K., Ragnarsson, R., Sigurgeirsson, B., Skuladottir, H., Gudbjartsson, T., Isaksson, H. J., et al. (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat. Genet.* **41**, 221–7.

Särndal, C., Swensson, B. & Wretman, J. (1992). *Model-Assisted Inference in Survey Sampling*. New York: Springer-Verlag.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–42.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W. V., Spitz, M. R., Eisen, T., Amos, C. I. & Houlston, R. S. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407–9.

Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D. W., Eyre, S., Freedman, B. I., Friedman, D. J., Field, J. K., Groop, L., Haugen, A., Heinrich, J., Henderson, B. E., Hicks, P. J., Hocking, L. J., Kolonel, L. N., et al. (2012). Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* **8**: e1003032 doi:10.1371/journal.pgen.1003032.

[*Received May* 2013. *Revised August* 2014]