STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2756–2783
Published online 18 September 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3044

Accounting for error due to misclassification of exposures in case–control studies of gene–environment interaction

Li Zhang[1], Bhramar Mukherjee[2,*,†], Malay Ghosh[3], Stephen Gruber[4] and Victor Moreno[5,6]

[1]*Department of Quantitative Health Sciences, The Cleveland Clinic Foundation, Cleveland, OH-44195, U.S.A.*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, MI-48109, U.S.A.*
[3]*Department of Statistics, University of Florida, Gainesville, FL-32611, U.S.A.*
[4]*Department of Internal Medicine, Epidemiology and Human Genetics, University of Michigan, Ann Arbor, MI-48109, U.S.A.*
[5]*Department of Internal Medicine and Epidemiology, University of Michigan, Ann Arbor, MI-48109, U.S.A.*
[6]*IDIBELL, Catalan Institute of Oncology, L'Hospitalet Barcelona, Spain*

SUMMARY

We consider analysis of data from an unmatched case–control study design with a binary genetic factor and a binary environmental exposure when both genetic and environmental exposures could be potentially misclassified. We devise an estimation strategy that corrects for misclassification errors and also exploits the gene–environment independence assumption. The proposed corrected point estimates and confidence intervals for misclassified data reduce back to standard analytical forms as the misclassification error rates go to zero. We illustrate the methods by simulating unmatched case–control data sets under varying levels of disease–exposure association and with different degrees of misclassification. A real data set on a case–control study of colorectal cancer where a validation subsample is available for assessing genotyping error is used to illustrate our methods. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: case-only method; gene–environment independence; sensitivity; specificity

1. INTRODUCTION

Measurement error in exposure assessment is one of the major sources of bias in epidemiological studies. When ignored, even small errors in exposure assessment can result in biased point and

*Correspondence to: Bhramar Mukherjee, Department of Biostatistics, University of Michigan, Ann Arbor, MI-48109, U.S.A.
†E-mail: bhramar@umich.edu

Contract/grant sponsor: NSA; contract/grant number: H98230-06-1-0033
Contract/grant sponsor: Spanish Secretara de Estado de Universidades e Investigacin, Ministerio de Educacin y Ciencia

Received 6 March 2007
Accepted 10 July 2007

Copyright © 2007 John Wiley & Sons, Ltd.

interval estimates of the parameters and invalidate P-values of hypotheses tests. Widespread existence of exposure measurement and misclassification errors in epidemiological research may explain much of the inconsistent and inconclusive results currently reported in the literature. Various statistical techniques have been developed to correct for exposure measurement errors in epidemiological studies with accompanying validation substudies [1–3].

Bashir and Duffy [4] provided a general review of epidemiological methods for dealing specifically with measurement error and misclassification. Gustafson [5] presented a unified approach to characterize the consequences of ignoring mismeasurement on resulting indicators of exposure–disease association and demonstrated the use of Bayesian methods to adjust for mismeasurement. Rice and Holmans [6] obtained analytical formulae for correcting risks due to a single genetic factor in terms of the genotyping error probabilities for analyzing unmatched case–control studies. Later, Rice [7] proposed a full-likelihood approach to obtain estimates and confidence intervals for the parameters of interest in the presence of misclassification of a binary exposure in a matched case–control study. However, many of the above discussions on the effects of misclassification of exposure in genetic epidemiological studies have focused on the impact on the relative risk and/or sample size in studies of just a single factor. In contrast, less attention has been given to the effect of misclassification on the assessment of interactions between two or more factors. There is a significant volume of literature for handling missingness and measurement error in unmatched and matched case–control studies [8–13]. However, our focus remains on the particular context of studies of gene–environment interaction.

One of the major goals in many recent epidemiological studies has been to investigate the effect of genes on a disease, in combination with environmental exposures. In case–control studies of gene–environment association with disease, when genetic and environmental exposures can be assumed to be independent in the underlying population, one may exploit the independence in order to derive more efficient estimation techniques than the traditional logistic regression analysis [14–16]. Garcia-Closas *et al.* [17] showed that, under a set of conditions often satisfied in studies of gene–environment interactions, both differential and non-differential misclassifications of a binary environmental factor attenuate the multiplicative interaction effect towards the null value. Garcia-Closas *et al.* [18] proposed a simple approach to assess the impact of misclassification on bias in the estimation of multiplicative or additive interactions and on sample size requirements. They pointed out that, under misclassification of exposures, increased sample size is needed to attain the same power to detect the attenuated interaction. The focus of Garcia-Closas *et al.* [18] was primarily on study design issues under misclassification, and the authors did not propose corrected estimates of the parameters of interest, or inferential adjustments, if in fact misclassification is present in the data.

Recently, Cheng [19] proposed an innovative conditional likelihood-based approach to adjust for bias caused by genotyping errors in case-only studies by using the information from an internal validation study, obtained by genotyping a randomly sampled set of individuals twice. One of the major criticisms of a case-only study is the possible bias in the estimates when gene–environment independence assumption is violated. Conditioning on covariates which may introduce non-independence between genetic (G) and environmental (E) factors is viewed as a potential remedy to prevent against such biases. Cheng [19, 20] used conditional independence of G and E instead of marginal independence under a case-only design, introducing adjustments for the interaction odds ratio estimate in the presence of genotyping error [19]. However, no corrected estimates of main effects of G and E could be obtained by using case-only data. Moreover, a possible misclassification of E is not considered in [19]. While the case-only design considered

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

in Cheng [19] requires rare disease assumption, in the current paper, we consider the situation when the disease is not rare but knowledge regarding the marginal prevalence of the disease is available.

In this paper, we describe a relatively simple approach to adjust the estimation of the parameters of interest in case–control studies of G–E interaction in the presence of misclassification in both G and E. The proposed method exploits the G–E independence assumption and obtains corrected parameter estimates for all parameters of interest, and not just the interaction odds ratio. We consider an unmatched case–control setup, adapt and extend the work of Rice and Holmans [6] to the situation when one has a binary G and a binary E, both of which are potentially subject to misclassification, with the additional constraint that the joint distribution of G and E satisfies the assumption of independence.

For a single biallelic locus, genetic exposure has inherently three instead of two levels, and in some cases it may be worthwhile to use the full scale of data, especially when there is uncertainty about the genetic susceptibility model [19]. Our proposed approach can be easily extended to the corresponding 2×6 table where one would consider genotype data recorded as 0, 1 or 2 depending on the number of copies of the variant allele. We have indicated an outline of this extension in Appendix A.3. However, as the basic theory and numerical findings remain fairly similar in the current paper, we focus mainly on the 2×4 table. The 2×4 table is also used to represent a dominant or a recessive genetic susceptibility model which is fairly common in practice. Botto and Khury [21] present many reasons to consider the 2×4 table as a pivotal quantity in the analysis of G–E interactions.

In Section 2, we start with the standard formulation in terms of odds ratios of a 2×4 table. We then describe maximum likelihood (ML) estimation under the G–E independence assumption and obtain maximum likelihood estimations (MLEs) under this additional restriction. We first make a rare disease assumption, in which case we can obtain a closed-form expression for the MLEs and their asymptotic variances. We point out that the estimate of the G–E interaction parameter obtained by this approach, as expected, is exactly identical to the estimate obtained by the popular case-only approach [14]. By using data on both cases and controls, in addition to an efficient estimate of the interaction odds ratio, we obtain estimates of the main effects due to G and E as in the constrained ML approach of [15] for a log-linear model. With knowledge of the marginal prevalence of the disease in the population ($P(D=1)$), we can relax the rare disease assumption. In the latter situation, we can also obtain the constrained MLEs. Although the corresponding score equations do not have explicit closed-form solutions, numerical evaluation is extremely straightforward.

After this preliminary formulation with a perfectly measured data set, we delve into the issue of adjusting the estimates in the presence of misclassification. We first consider the situation with fixed values of sensitivity and specificity parameters of the measurement process. In the presence of misclassification, based on the sensitivity and specificity of the measuring instruments for genetic and environmental factors, we adjust the MLEs for bias due to misclassification. Corrected test statistics and confidence intervals are formulated as in any standard likelihood-based inference using the asymptotic distribution of the MLE, once the adjustments are made. In fact, as misclassification error rates go to zero, the estimates reduce to the standard MLEs for a perfectly recorded data set, had there been one. We also provide comparisons for the proposed methods in terms of coverage probabilities and power. Corrections for the interaction odds ratio under a case-only design follow directly and are discussed in Section 2.3. In Section 2.4, we briefly describe how to estimate the sensitivity and specificity parameters, which are typically unknown, based on a validation substudy. We review several options to obtain the estimates of the error rates based on a validation study, using both frequentist and Bayesian recipes [22]. Simulation studies in Section 3 show that our corrected inference substantially reduces bias when compared with the unadjusted inference based on misclassified cell counts. Section 4 contains the analysis of a real case–control study on colorectal cancer where data on a 'gold standard' measurement are available from a validation substudy. Finally, Section 5 contains a concluding discussion, whereas proofs and detailed calculations are relegated to the Appendix.

2. THE 2×4 TABLE

We consider unmatched case–control studies with a binary genetic factor G and a binary environmental exposure E, which take values 1 for susceptible (exposed in the case of E) and 0 for non-susceptible (unexposed in the case of E) subjects. Let D denote the disease status, where $D=1$ denotes affected, and $D=0$ denotes unaffected individuals. Using the same notation as in [18], the odds ratio OR_{eg} measures the association between disease and the environmental and genetic factors. Relative to subjects not exposed to the environmental or genetic factor ($E=0$ and $G=0$ are treated as the baseline categories), we define the following odds ratios: OR_{10} denotes the odds ratio for non-susceptible subjects exposed to the environmental factor; OR_{01} denotes the odds ratio for susceptible subjects not exposed to the environmental factor; and OR_{11} denotes the odds ratio for susceptible subjects exposed to the environmental factor. Therefore, $\psi=\mathrm{OR}_{11}/(\mathrm{OR}_{10}\,\mathrm{OR}_{01})$ is the multiplicative interaction parameter.

2.1. *MLE under G–E independence assumption*

Table I presents a general format of the data that we are considering. In the absence of misclassification, we can assume that the cell frequencies in the control and case populations, follow independent multinomial distributions namely $\mathbf{r}_0\sim Mn(n_0,\mathbf{p}_0)$ and $\mathbf{r}_1\sim Mn(n_1,\mathbf{p}_1)$, where n_0 and n_1 are fixed, and $\mathbf{r}_0=(r_{01},r_{02},r_{03},r_{04})$, $\mathbf{r}_1=(r_{11},r_{12},r_{13},r_{14})$, $\mathbf{p}_0=(p_{01},p_{02},p_{03},p_{04}=1-p_{01}-p_{02}-p_{03})$ and $\mathbf{p}_1=(p_{11},p_{12},p_{13},p_{14}=1-p_{11}-p_{12}-p_{13})$. By the definition of the odds ratios, we have $\mathrm{OR}_{10}=p_{01}p_{12}/(p_{02}p_{11})$, $\mathrm{OR}_{01}=p_{01}p_{13}/(p_{03}p_{11})$, $\mathrm{OR}_{11}=p_{01}p_{14}/(p_{04}p_{11})$ and $\psi=p_{02}p_{03}p_{11}p_{14}/(p_{01}p_{04}p_{12}p_{13})$. Thus, we obtain the case probabilities parameterized in terms of the relevant odds ratios and the control probabilities as $p_{11}=p_{01}/p$, $p_{12}=p_{02}/p\cdot\mathrm{OR}_{10}$, $p_{13}=p_{03}/p\cdot\mathrm{OR}_{01}$ and $p_{14}=p_{04}/p\cdot\mathrm{OR}_{10}\cdot\mathrm{OR}_{01}\cdot\psi$, where $p=p_{01}+p_{02}\cdot\mathrm{OR}_{10}+p_{03}\cdot\mathrm{OR}_{01}+p_{04}\cdot\mathrm{OR}_{10}\cdot\mathrm{OR}_{01}\cdot\psi$. The corresponding multinomial likelihood is given by

$$L_1=L(\mathrm{OR}_{10},\mathrm{OR}_{01},\psi,p_{01},p_{02},p_{03}|\mathbf{r}_0,\mathbf{r}_1)=\prod_{d=0}^{1}\prod_{j=1}^{4}p_{dj}^{r_{dj}} \tag{1}$$

Note that the parameterization in terms of p_{01}, p_{02} and p_{03} imposes no other restrictions except that they lead to valid probability distributions (all positive and summation less than 1). Similarly, the odds ratios are required to be positive. We can easily maximize the likelihood (1) and obtain the MLEs of the parameters of interest and their estimated asymptotic variance ($\widehat{\mathrm{AVAR}}$) as in Table II under the column of *unconstrained model*. The MLEs of the cell probabilities are simply given by $\hat{p}_{dj}=r_{dj}/n_d$, $d=0,1$, $j=1,\ldots,4$.

Let us now describe how the estimation changes with the additional constraint of G–E independence in the source population. We first investigate the estimates under a rare disease assumption,

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table I. Data for an unmatched case–control study with a binary genetic factor and a binary environmental exposure.

	$G=0$		$G=1$		
	$E=0$	$E=1$	$E=0$	$E=1$	Total
j	1	2	3	4	
$D=0$	r_{01}	r_{02}	r_{03}	r_{04}	n_0
$D=1$	r_{11}	r_{12}	r_{13}	r_{14}	n_1

Table II. The MLEs of the odds ratios and their estimated asymptotic variances in terms of observed counts r_{dj} for both the traditional unconstrained model and the model under G–E independence and rare disease in the absence of misclassification.

Parameters		Unconstrained model	G–E independence and rare disease
$\log(\mathrm{OR}_{10})$	MLE	$\log(r_{01}r_{12})-\log(r_{02}r_{11})$	$\log(r_{12}(r_{01}+r_{03}))-\log(r_{11}(r_{02}+r_{04}))$
	$\widehat{\mathrm{AVAR}}$	$\frac{1}{r_{01}}+\frac{1}{r_{02}}+\frac{1}{r_{11}}+\frac{1}{r_{12}}$	$\frac{1}{r_{01}+r_{03}}+\frac{1}{r_{02}+r_{04}}+\frac{1}{r_{11}}+\frac{1}{r_{12}}$
$\log(\mathrm{OR}_{01})$	MLE	$\log(r_{01}r_{13})-\log(r_{03}r_{11})$	$\log(r_{13}(r_{01}+r_{02}))-\log(r_{11}(r_{03}+r_{04}))$
	$\widehat{\mathrm{AVAR}}$	$\frac{1}{r_{01}}+\frac{1}{r_{03}}+\frac{1}{r_{11}}+\frac{1}{r_{13}}$	$\frac{1}{r_{01}+r_{02}}+\frac{1}{r_{03}+r_{04}}+\frac{1}{r_{11}}+\frac{1}{r_{13}}$
$\log(\psi)$	MLE	$\log(r_{02}r_{03}r_{11}r_{14})-\log(r_{01}r_{04}r_{12}r_{13})$	$\log(r_{11}r_{14})-\log(r_{12}r_{13})$
	$\widehat{\mathrm{AVAR}}$	$\sum_{d=0}^{1}\sum_{j=1}^{4}(1/r_{dj})$	$\sum_{j=1}^{4}(1/r_{1j})$

which is routinely made in epidemiological studies. The assumption of G–E independence in the source population, $P(G,E)=P(G)P(E)$, in conjunction with the rare disease assumption, implies that G–E independence holds in the control population, i.e. $P(G,E|D=0)=P(G|D=0)P(E|D=0)$. This adds an additional restriction on p_{01}, p_{02} and p_{03}, namely

$$p_{01}(1-p_{01}-p_{02}-p_{03})=p_{02}p_{03} \tag{2}$$

With this additional restriction, maximizing the likelihood (1) will not provide the same estimates as in the traditional unconstrained model. The MLEs and their $\widehat{\mathrm{AVAR}}$ in this restricted parameter space are presented in Table II under the column *G–E independence and rare disease*. The constrained ML equations and their solutions which lead to this column in Table II are presented in Appendix A.1. Note that the asymptotic variance under the unconstrained model is always larger than that under the constrained model. Gain in efficiency in the MLEs obtained from the retrospective likelihood when constraints on the exposure distribution (such as G–E independence or Hardy–Weinberg equilibrium) are exploited has been noted in several recent papers [16, 23, 24].

If the disease prevalence $P(D=1)=\pi$ in the source population is known, we can relax the rare disease assumption by expressing the G–E independence as the following:

$$\begin{aligned}P(G=g)P(E=e)&=P(G=g,E=e)\\&=P(G=g,E=e|D=0)P(D=0)+P(G=g,E=e|D=1)P(D=1)\end{aligned} \tag{3}$$

Copyright © 2007 John Wiley & Sons, Ltd.

where $g, e = 0, 1$. Therefore, instead of the restriction as in (2), we have the following restriction on p_{01}, p_{02} and p_{03}:

$$\begin{aligned} f &= (1-\pi)p_{04} + \pi \mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi p_{04}/p \\ &\quad - [(1-\pi)(p_{02}+p_{04}) + \pi(\mathrm{OR}_{10}\,p_{02} + \mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi p_{04})/p] \\ &\quad \times [(1-\pi)(p_{03}+p_{04}) + \pi(\mathrm{OR}_{01}\,p_{03} + \mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi p_{04})/p] = 0 \end{aligned} \tag{4}$$

The details of obtaining (4) are deferred to Appendix A.2. With this additional restriction, maximizing the likelihood in (1) will not provide the same estimates as under the rare disease assumption. In fact, the solutions to the ML equations cannot be written in closed form. However, we can obtain the restricted MLEs by the usual Newton–Raphson algorithm and obtain the estimated asymptotic variance–covariance matrix by the inverse of the observed information matrix. The observed information matrix is constructed by taking the second derivative of the log-likelihood with respect to the parameters and evaluating them at the MLEs of the parameters.

Because of the skewness in the sampling distribution of the estimated odds ratios, statistical inference for the odds ratio parameters (denoted by a generic symbol θ) uses an alternative but equivalent measure: its natural logarithm, $\log(\hat{\theta})$. By simple use of the delta method, the large-sample distribution of $\log(\hat{\theta})$ is approximately normal, i.e. $\log(\hat{\theta}) \sim \mathrm{N}(\log(\theta), \mathrm{AVAR}(\log\hat{\theta}))$, where $\hat{\theta}$ is the MLE of θ, and the estimated asymptotic variance of $\log(\hat{\theta})$ is obtained from the observed Fisher information. Standard z-tests and confidence intervals for the log-scale parameters are constructed on the basis of the above asymptotic normality.

Remark 1

It is well known that, in a multinomial setup, the expected cell counts, namely $E_{\mathbf{p}_d}[\mathbf{r}_d]$, are simply equal to $n_d\mathbf{p}_d$, where $E_{\mathbf{p}_d}[\mathbf{r}_d]$ represents the row vector of expected cell counts corresponding to $D = d, d = 0, 1$, and $\mathbf{p}_d$ denotes the true probability vector. Then, the vector of *estimated* expected cell frequencies, denoted by $\tilde{\mathbf{r}}_d$, is given by $\tilde{\mathbf{r}}_d = E_{\mathbf{p}_d}[\mathbf{r}_d]|_{\mathbf{p}_d=\hat{\mathbf{p}}_d} = n_d\hat{\mathbf{p}}_d$, (i.e. the expected frequencies evaluated at the MLEs of the model parameters). For example, for the usual multinomial model, without any restrictions on the exposure space, the vector of estimated expected cell frequencies matches exactly with the observed frequencies, that is $\tilde{\mathbf{r}}_d = \mathbf{r}_d$ (as $\hat{\mathbf{p}}_d = \mathbf{r}_d/n_d$), where $\mathbf{r}_d$ is the vector of observed frequencies.

Under G–E independence and rare disease assumptions (denoted by the superscript IR below, to distinguish from the other models), from Appendix A.1, we note that the MLEs for $\mathbf{p}$ are

$$\begin{aligned} &\hat{p}_{01}^{IR} = \frac{(r_{01}+r_{03})(r_{01}+r_{02})}{n_0^2}, \quad \hat{p}_{02}^{IR} = \frac{(r_{01}+r_{02})(r_{02}+r_{04})}{n_0^2} \\ &\hat{p}_{03}^{IR} = \frac{(r_{01}+r_{03})(r_{03}+r_{04})}{n_0^2}, \quad \hat{p}_{04}^{IR} = \frac{(r_{02}+r_{04})(r_{03}+r_{04})}{n_0^2} \\ &\hat{p}_{1j}^{IR} = \frac{r_{1j}}{n_1}, \quad j = 1, 2, 3, 4 \end{aligned} \tag{5}$$

and thus the estimated expected frequencies are obtained simply by $\tilde{\mathbf{r}}_d^{IR} = n_d\,\hat{\mathbf{p}}_d^{IR}$.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Remark 2
One can obtain the estimates of the marginal odds ratios for G and E by using the estimates of OR_{10}, OR_{01} and the interaction effect ψ, as well as the cell probabilities in the control population, $\mathbf{p}_0$. Define the genetic and environmental marginal odds ratios OR_G and OR_E as the following:

$$\mathrm{OR}_E = \frac{P(D=1|E=1)P(D=0|E=0)}{P(D=0|E=1)P(D=1|E=0)}$$
$$\mathrm{OR}_G = \frac{P(D=1|G=1)P(D=0|G=0)}{P(D=0|G=1)P(D=1|G=0)}$$

Thus, one is able to estimate OR_G and OR_E by using the following identities:

$$\mathrm{OR}_E = \left\{\frac{p_{01}+p_{03}}{p_{02}+p_{04}}\right\}\left\{\frac{p_{02}\,\mathrm{OR}_{10}+p_{04}\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi}{p_{01}+p_{03}\,\mathrm{OR}_{01}}\right\}$$
$$\mathrm{OR}_G = \left\{\frac{p_{01}+p_{02}}{p_{03}+p_{04}}\right\}\left\{\frac{p_{03}\,\mathrm{OR}_{01}+p_{04}\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi}{p_{01}+p_{02}\,\mathrm{OR}_{10}}\right\}$$

Under the G–E independence and rare disease assumptions, we have $p_{01}p_{04}=p_{02}p_{03}$. Furthermore, $P(E|D=0)\approx P(E)$ and $P(G|D=0)\approx P(G)$; thus, one can estimate OR_G and OR_E by using

$$\mathrm{OR}_E = \frac{(1-P(G=1))\,\mathrm{OR}_{10}+P(G=1)\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi}{(1-P(G=1))+P(G=1)\,\mathrm{OR}_{01}}$$
$$\mathrm{OR}_G = \frac{(1-P(E=1))\,\mathrm{OR}_{01}+P(E=1)\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi}{(1-P(E=1))+P(E=1)\,\mathrm{OR}_{10}}$$

2.2. MLE in the presence of misclassification

In this section, we introduce the effects of misclassification into the estimation framework. Our model for misclassified data is based on the assumption that some perfectly classified 'true' case–control data exist, where the true underlying cell probabilities follow the same model as $\mathbf{p}_d$ discussed above. Following the 'star' notation of [6], we let the superscript asterisk denote the true parameters for the true data model as well as the perfectly measured exposure variables. Let sp_{dG} (se_{dG}) and sp_{dE} (se_{dE}) denote specificity (sensitivity) of G and E with disease status d, respectively, where sensitivity $=P$(observed exposed | truly exposed) and specificity $=P$(observed unexposed | truly unexposed); hence, $se_{dG}=P(G=1|G^*=1, D=d)$, $se_{dE}=P(E=1|E^*=1, D=d)$, $sp_{dG}=P(G=0|G^*=0, D=d)$ and $sp_{dE}=P(G=0|G^*=0, D=d)$. Applying a classical error structure, all subjects are assumed to have the same probability of the observed exposure, conditional on their case/control status and true exposure. We then have the following two results.

Result 1
Assuming that given the disease status $d(=0, 1)$ and the true exposure status of G and E the observed exposure statuses of G and E are independent, then

$$\begin{pmatrix} p_{d1} & p_{d2} \\ p_{d3} & p_{d4} \end{pmatrix} = \mathbf{A}_d \begin{pmatrix} p^*_{d1} & p^*_{d2} \\ p^*_{d3} & p^*_{d4} \end{pmatrix} \mathbf{B}_d \qquad (6)$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

where

$$\mathbf{A}_d=\begin{pmatrix} sp_{dG} & 1-se_{dG} \\ 1-sp_{dG} & se_{dG} \end{pmatrix} \quad \text{and} \quad \mathbf{B}_d=\begin{pmatrix} sp_{dE} & 1-sp_{dE} \\ 1-se_{dE} & se_{dE} \end{pmatrix}$$

Proof

$$\begin{aligned} &P(G,E|D=d)\\ &=\sum_{g=0}^{1}\sum_{e=0}^{1}P(G,E|D=d,G^*=g,E^*=e)P(G^*=g,E^*=e|D=d)\\ &=\sum_{g=0}^{1}\sum_{e=0}^{1}P(G|D=d,G^*=g,E^*=e)P(E|D=d,G^*=g,E^*=e)P(G^*=g,E^*=e|D=d)\\ &=\sum_{g=0}^{1}\sum_{e=0}^{1}P(G|D=d,G^*=g)P(E|D=d,E^*=e)P(G^*=g,E^*=e|D=d)\end{aligned}$$

p_{dj} as defined in Table I denotes the cell probabilities of the jth $(j=1,\ldots,4)$ (G,E) configuration given the disease status $d=0,1$. Note that the second equality is not a result of the G–E independence assumption, but, given the disease status $d(=0,1)$ and the true exposure statuses of G and E, the observed exposure statuses of G and E are independent. Result 1 holds for all three models discussed in the previous section. Therefore, if the observed data come from a common multinomial distribution with cell probabilities p_{dj}, then we can write down the likelihood (1) in terms of the true, 'starred' parameters. We simply write the p_{dj}'s in terms of a linear function of the true parameters p^*_{dj} as defined by Result 1 and maximize the following multinomial likelihood in terms of the underlying true or starred parameters

$$L_2=L(\mathrm{OR}^*_{10},\mathrm{OR}^*_{01},\psi^*,p^*_{01},p^*_{02},p^*_{03}|\mathbf{r}_0,\mathbf{r}_1)=\prod_{d=0}^{1}\prod_{j=1}^{4}\{p_{dj}(\mathbf{p}^*_d)\}^{r_{dj}} \tag{7}$$

where $p_{dj}(\mathbf{p}^*_d)$ denotes the linear transformation defined in (6); essentially we are replacing the p_{dj} in the original likelihood by a function of the underlying true parameters as described in Result 1. Thus, by maximizing the likelihood (7), which now includes the effect of misclassification through the linear transformation on the parameters with the correction matrices $\mathbf{A}_d$ and $\mathbf{B}_d$, we can now obtain the MLEs of the starred parameters, denoted by $\hat{\mathbf{p}}^*_d$.

As indicated in Remark 1, the vector of estimated expected cell counts under the multinomial model is given by $\tilde{\mathbf{r}}_d=n_d\ \hat{\mathbf{p}}_d$. Thus, for the estimation with the starred parameters, the vector of estimated expected cell counts under the true data model is $\mathbf{r}^*_d=n_d\ \hat{\mathbf{p}}^*_d$. Note that by the invariance property of the MLE, Result 1 holds when the parameters $\mathbf{p}_0$ and $\mathbf{p}_1$ are replaced with the MLEs for the perfectly classified data model and the misclassified data model. Thus, by inverting Result 1 as in (6), replacing the parameters with the MLEs, we have

$$\begin{pmatrix} \hat{p}^*_{d1} & \hat{p}^*_{d2} \\ \hat{p}^*_{d3} & \hat{p}^*_{d4} \end{pmatrix}=\mathbf{A}_d^{-1}\begin{pmatrix} \hat{p}_{d1} & \hat{p}_{d2} \\ \hat{p}_{d3} & \hat{p}_{d4} \end{pmatrix}\mathbf{B}_d^{-1}=\frac{1}{n_d}\mathbf{A}_d^{-1}\begin{pmatrix} \tilde{r}_{d1} & \tilde{r}_{d2} \\ \tilde{r}_{d3} & \tilde{r}_{d4} \end{pmatrix}\mathbf{B}_d^{-1}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table III. The MLEs of the true odds ratios in terms of estimated starred expected counts r^*_{dj} for the traditional unconstrained model (Model 1) and ${r^*_{dj}}^{IR}$ for the model under *G–E* independence and rare disease assumptions (Model 2) in the presence of misclassification.

Parameters	Model 1	Model 2
$\log(\mathrm{OR}^*_{10})$	$\log(r^*_{01}r^*_{12})-\log(r^*_{02}r^*_{11})$	$\log(r^{*IR}_{01}r^{*IR}_{12})-\log(r^{*IR}_{02}r^{*IR}_{11})$
$\log(\mathrm{OR}^*_{01})$	$\log(r^*_{01}r^*_{13})-\log(r^*_{03}r^*_{11})$	$\log(r^{*IR}_{01}r^{*IR}_{13})-\log(r^{*IR}_{03}r^{*IR}_{11})$
$\log(\psi^*)$	$\log(r^*_{02}r^*_{03}r^*_{11}r^*_{14})-\log(r^*_{01}r^*_{04}r^*_{12}r^*_{13})$	$\log(r^{*IR}_{11}r^{*IR}_{14})-\log(r^{*IR}_{12}r^{*IR}_{13})$

This immediately leads to the following relationship between estimated expected cell counts for the true data and the misclassified data. □

Result 2

$$\begin{pmatrix} r^*_{d1} & r^*_{d2} \\ r^*_{d3} & r^*_{d4} \end{pmatrix} = \mathbf{A}_d^{-1} \begin{pmatrix} \tilde{r}_{d1} & \tilde{r}_{d2} \\ \tilde{r}_{d3} & \tilde{r}_{d4} \end{pmatrix} \mathbf{B}_d^{-1} \tag{8}$$

In fact, the result is true for the vector of expected cell counts involving the unknown parameters, not only the sample estimates, as is obvious from the above discussion.

Thus, for the traditional multinomial model and the model under the *G–E* independence and rare disease assumptions, the MLEs of the true starred parameters of interest have a closed-form expression in terms of the estimated starred expected cell counts $\mathbf{r}^*_d$, which are shown in Table III. To obtain $\mathbf{r}^*_d$, we simply obtain the MLEs $\hat{\mathbf{p}}^*_d$ under different models and multiply by n_d. Note that the MLEs $\hat{\mathbf{p}}^*_d$ are also easily obtained by using the transformation in Result 1 and the ML estimation of $\mathbf{p}_d$ as discussed in Section 2.1 under different model assumptions. The MLEs $\hat{\mathbf{p}}^*_d$ turn out to be different functions of the observed cell counts $\mathbf{r}_d$, sensitivity and specificity parameters, the form of the function depending on the model assumptions. Therefore, $\mathbf{r}^*_d$ under different assumptions or constraints on the parameters might be different (we use the superscript IR to denote under *G–E* independence and rare disease assumptions to distinguish it from other models) as the MLEs $\hat{\mathbf{p}}^*_d$ (and $\hat{\mathbf{p}}_d$) are different across models with different assumptions. (We refer to the discussion comparing the usual multinomial model, and the model with rare disease and *G–E* independence in Section 2.1). This simply means that we can apply the corrected counts instead of the observed counts $\mathbf{r}_d$ to the estimates obtained in Table II, which will lead to the exactly same estimates as described in Table III. We emphasize that these estimators in Table III are strictly valid as MLEs only when they lie within the constrained parameter space. When the positivity constraints on the OR^*_{eg}'s or the probability constraints on the $\mathbf{p}^*_d$'s are violated (e.g. when very small values of sensitivity or specificity are used, corresponding to huge misclassification rates), the constrained MLEs would be on the boundary of the parameter space. We should then maximize the likelihood (7) directly with respect to the true parameters subject to the constraints, instead of transforming the observed MLE. However, as such estimates are indicative of extreme misclassification, or too small a sample, we might as well treat these estimates with some caution.

Construction of the confidence intervals follows in exactly the same way as for the perfectly classified data, with the standard error estimates obtained from the inverse of the information matrix of L_2 evaluated at the MLEs.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

We can also observe the behavior of the estimators from Table III as the misclassification error rates go to 0 by Taylor series expansions of these estimators. Define the errors as $\varepsilon^p_{dG}=1-sp_{dG}$, $\varepsilon^e_{dG}=1-se_{dG}$, $\varepsilon^p_{dE}=1-sp_{dE}$ and $\varepsilon^e_{dE}=1-se_{dE}$. Expanding the log-scale estimators $\hat{\psi}^*$ and $\hat{\psi}^{*IR}$ of the interaction parameter around $\varepsilon^p_{dG}=\varepsilon^e_{dG}=\varepsilon^p_{dE}=\varepsilon^e_{dE}=0$, we see that

$$\begin{aligned}
&\log(\hat{\psi}^*)\\
&\quad=\log\left(\frac{r^*_{02}r^*_{03}r^*_{11}r^*_{14}}{r^*_{01}r^*_{04}r^*_{12}r^*_{13}}\right)\\
&\quad=\log\left(\frac{r_{02}r_{03}r_{11}r_{14}}{r_{01}r_{04}r_{12}r_{13}}\right)+\frac{(r_{01}r_{04}-r_{02}r_{03})(r_{03}r_{04}\varepsilon^e_{0G}+r_{02}r_{04}\varepsilon^e_{0E}+r_{01}r_{02}\varepsilon^p_{0G}+r_{01}r_{03}\varepsilon^p_{0E})}{r_{01}r_{02}r_{03}r_{04}}\\
&\qquad+\frac{(r_{11}r_{14}-r_{12}r_{13})(r_{13}r_{14}\varepsilon^e_{1G}+r_{12}r_{14}\varepsilon^e_{1E}+r_{11}r_{12}\varepsilon^p_{1G}+r_{11}r_{13}\varepsilon^p_{1E})}{r_{11}r_{12}r_{13}r_{14}}+\text{higher order terms}\\
&\qquad\log(\hat{\psi}^{*IR})=\log\left(\frac{r^{*IR}_{11}r^{*IR}_{14}}{r^{*IR}_{12}r^{*IR}_{13}}\right)\\
&\quad=\log\left(\frac{r_{11}r_{14}}{r_{12}r_{13}}\right)+\frac{(r_{11}r_{14}-r_{12}r_{13})(r_{13}r_{14}\varepsilon^e_{1G}+r_{12}r_{14}\varepsilon^e_{1E}+r_{11}r_{12}\varepsilon^p_{1G}+r_{11}r_{13}\varepsilon^p_{1E})}{r_{11}r_{12}r_{13}r_{14}}\\
&\qquad+\text{higher order terms}
\end{aligned}\tag{9}$$

Up to a first-order approximation, the estimator reduces to the normal, perfect-data estimate of the interaction odds ratio described in Table II as the error rates go to 0. The first-order terms suggest that using a good approximation to errors may give better estimates than those by simply ignoring misclassification, i.e. setting the errors equal to 0. The expressions also suggest that the misclassification probably affects the unconstrained estimate of ψ more as there is contribution from two such first-order error terms.

2.3. *Case-only method with possible misclassification*

The case-only method [14] is a popular method to estimate the multiplicative G–E interaction parameter ψ, where, under the rare disease and G–E independence assumptions, the odds ratio of G for exposed versus unexposed subjects among the cases only provides an efficient estimate of ψ. The data used are as shown in the second row of Table I, ignoring the control data in the first row.

In the absence of misclassification, data from the case population form a multinomial distribution, $\mathbf{r}_1\sim Mn(n_1,\mathbf{p}_1)$, where n_1 is fixed. The interaction parameter (denoted here as ψ^{CO}) is obtained as the odds ratio between G and E among the case population, i.e. $\psi^{\text{CO}}=p_{11}p_{14}/(p_{12}p_{13})$. Together with $\sum_{j=1}^4 p_{1j}=1$, we have the following restrictions for $\mathbf{p}_1$:

$$p_{13}=\frac{p_{11}(1-p_{11}-p_{12})}{p_{11}+p_{12}\psi^{\text{CO}}}\quad\text{and}\quad p_{14}=\frac{p_{12}\psi^{\text{CO}}(1-p_{11}-p_{12})}{p_{11}+p_{12}\psi^{\text{CO}}}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

The corresponding likelihood for the case-only method is thus $L_{\mathrm{CO}} = L(\psi^{\mathrm{CO}}, p_{11}, p_{12}|\mathbf{r}_1) = \prod_{j=1}^{4} p_{1j}^{r_{1j}}$, and the MLE of the interaction parameter ψ^{CO} is $\hat{\psi}^{\mathrm{CO}} = r_{11}r_{14}/(r_{12}r_{13})$ with variance $\{\hat{\psi}^{\mathrm{CO}}\}^{-2}(\sum_{j=1}^{4} 1/r_{1j})$.

Both Results 1 and 2 hold for $d=1$ as well; therefore, estimating the true parameters in the presence of misclassification is straightforward by writing the likelihood in terms of the true parameters $L^*_{\mathrm{CO}} = L(\psi^{*\mathrm{CO}}, p^*_{11}, p^*_{12}|\mathbf{r}_1) = \prod_{j=1}^{4} \{p_{1j}(\mathbf{p}^*_1)\}^{r_{1j}}$. Note that $\hat{p}_{1j} = r_{1j}/n_1$; hence, the MLE of the 'true' parameter $\psi^{*\mathrm{CO}}$ in terms of $\mathbf{r}_1^{*\mathrm{CO}}$ is $\hat{\psi}^{*\mathrm{CO}} = r_{11}^{*\mathrm{CO}} r_{14}^{*\mathrm{CO}}/(r_{12}^{*\mathrm{CO}} r_{13}^{*\mathrm{CO}})$, and $r_{1j}^{*\mathrm{CO}}$ can be obtained following Result 2, with $\tilde{r}_{1j} = n_1\hat{p}_{1j} = r_{1j}$. The variance estimators can again be estimated from the inverse of the information matrix of L^*_{CO} evaluated at the MLEs or by the technique as stated in Appendix A.1.

Remark 3

Note that the MLE of the interaction parameter and its variance obtained by the case-only method are exactly the same as those obtained in Section 2.1, where we also assume G–E independence and rare disease assumptions, but use both case and control data. This is true whether in the absence of misclassification or in the presence of misclassification (unadjusted or adjusted). This establishes yet another proof of the fact that, under G–E independence and rare disease assumptions, the MLE of the interaction odds ratio is exactly equal to the odds ratio of E on G for cases alone.

Remark 3 shows that our model with the G–E independence and rare disease assumptions can also obtain a highly efficient estimate of the interaction parameter ψ as in the case-only method. Moreover, our model is also able to estimate the main effects of genetic and environmental factors, which the case-only method cannot estimate (Umbach and Weinberg [15] established this for a more general log-linear model). As Clayton and McKeigue [25] pointed out, studies of gene–environment association with disease need to go beyond the mere estimation of the statistical interaction parameter ψ, and our study can estimate auxiliary parameters like the joint effects of interest without compromising on the efficiency of the estimate of the interaction parameter. An outline of extending our method to a 2 × 6 table, when genotype data are recorded into three levels, is presented in Appendix A.3.

2.4. *Validation studies when misclassification rates are unknown*

Instead of knowing the various misclassification rates perfectly, we shall briefly describe a commonly used strategy to estimate the misclassification rates from a validation substudy. In case the validation studies are independent of each other and of the main study, the full joint likelihood including the validation study is given by

$$L_{\mathrm{Full}} = L_2 \times L_{\mathrm{valid}} \tag{10}$$

where $L_2 = \prod_{d=0}^{1}\prod_{j=1}^{4} \{p_{dj}(\mathbf{p}^*_d)\}^{r_{dj}}$ as in (9). In our simulation studies we work with this setup. For illustration purposes, assume that there is no misclassification in G. We shall now assume that each of the four misclassification rates ($\boldsymbol{\tau} = (sp_{0E}, sp_{1E}, se_{0E}, se_{1E})$) of E has been estimated from a binomial validation study of 100 subjects, so that, for example, sp_{0E} is known only through a correctly classified sample of size (n_{0p}) drawn from a Bin(100, sp_{0E}) distribution. Let n_{1p}, n_{0e}, n_{1e} denote the sizes of correctly classified samples drawn from the corresponding binomial distributions when estimating $sp_{1E}, se_{0E}, se_{1E}$, respectively. Then the likelihood from the validation study

Copyright © 2007 John Wiley & Sons, Ltd.

is given by

$$L_{\text{valid}}(\boldsymbol{\tau}|n_{0p}, n_{1p}, n_{0e}, n_{1e}) = (sp_{0E})^{n_{0p}}(1-sp_{0E})^{100-n_{0p}}(sp_{1E})^{n_{1p}}(1-sp_{1E})^{100-n_{1p}} \\ \times (se_{0E})^{n_{0e}}(1-se_{0E})^{100-n_{0e}}(se_{1E})^{n_{1e}}(1-se_{1E})^{100-n_{1e}} \quad (11)$$

If we need to introduce misclassification in G, there will be four more such binomial probability contributions. The analyses of the previous section assumed known fixed misclassification rates $\boldsymbol{\tau}$; thus the likelihood had contribution only from L_2. Now we consider estimating $\boldsymbol{\tau}$ from the validation study by the following two ways:

(i) *A crude plug-in method*: Estimate $\boldsymbol{\tau}$ from L_{valid} *only* and plug in the estimates in L_2 . Then maximize L_2 exactly as in Section 2.2. This naive method ignores uncertainty in the obtained estimates of sensitivity and specificity parameters, and will lead to smaller coverage probabilities for the confidence intervals for the odds ratio parameters than the designated confidence levels.

(ii) *Joint estimation of all unknown parameters*: Obtain the estimates of OR^*_{eg} $(e, g=0, 1)$, p^*_{0j} $(j=1, 2, 3)$ and $\boldsymbol{\tau}$ simultaneously based on the full joint likelihood as in (10). There are several options to achieve this: (a) maximize the full likelihood (10) with respect to OR^*_{eg} $(e, g=0, 1)$, p^*_{0j} $(j=1, 2, 3)$ and $\boldsymbol{\tau}$ to obtain the MLEs; (b) take a full Bayesian approach by introducing priors on all parameters and use the Markov chain Monte Carlo to conduct posterior inference; (c) use the marginal likelihood approach where one integrates out $\boldsymbol{\tau}$ with respect to Uniform(0, 1) priors on each error rate parameter; and (d) maximize the integrated likelihood in terms of OR^*_{eg} $(e, g=0, 1)$ and p^*_{0j} $(j=1, 2, 3)$. Note that L_{valid} involves only $\boldsymbol{\tau}$; hence, the MLEs of OR^*_{eg} $(e, g=0, 1)$ and p^*_{0j} $(j=1, 2, 3)$ have the same formulations as when $\boldsymbol{\tau}$ is known (under the corresponding model assumptions), which are the functions of $\mathbf{r}_0$, $\mathbf{r}_1$ and $\boldsymbol{\tau}$, denoted as $\widehat{\text{OR}}^*_{eg}(\boldsymbol{\tau})$ and $\hat{p}^*_{0j}(\boldsymbol{\tau})$. Therefore, one can first estimate $\boldsymbol{\tau}$ based on the following profile likelihood:

$$\begin{aligned} &L_p(\widehat{\text{OR}}^*_{eg}(\boldsymbol{\tau}), \hat{p}^*_{0j}(\boldsymbol{\tau}), \boldsymbol{\tau}|\mathbf{r}_0, \mathbf{r}_1, n_{0p}, n_{1p}, n_{0e}, n_{1e}) \\ &\quad = L_2(\widehat{\text{OR}}^*_{eg}(\boldsymbol{\tau}), \hat{p}^*_{0j}(\boldsymbol{\tau})|\mathbf{r}_0, \mathbf{r}_1)L_{\text{valid}}(\boldsymbol{\tau}|n_{0p}, n_{1p}, n_{0e}, n_{1e}) \end{aligned} \quad (12)$$

The profile likelihood is maximized by the usual Newton–Raphson method to obtain the MLEs of $\boldsymbol{\tau}$. One can then plug in the estimates of $\boldsymbol{\tau}$ into $\widehat{\text{OR}}^*_{eg}(\boldsymbol{\tau})$ and $\hat{p}^*_{0j}(\boldsymbol{\tau})$ to obtain the estimates of OR^*_{eg} $(e, g=0, 1)$ and p^*_{0j} $(j=1, 2, 3)$. We present the results of the crude plug-in method and the last profile likelihood approach, although each of the four methods offers satisfactory solution when parameters do not lie on the boundaries. Since the crude plug-in method is easiest to implement and provides results fairly comparable to those obtained using methods based on maximizing the joint likelihood, we advocate using this method, knowing that the actual confidence intervals accounting for uncertainties in the estimation of sensitivity/specificity parameters are marginally wider than the ones obtained in most cases.

Remark 4

Rice and Holmans [6], Cheng [19] and Lai *et al.* [26] all considered validation data generated using the repeated measurement sampling method, which genotypes repeatedly a fraction of the sampled subjects with the same error-prone genotyping instrument to obtain estimates of misclassification probabilities. The assumption typically made in much of the literature is that genotyping errors are the same for cases and controls and are independent of one another. In our formulation, the

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

first assumption is relaxed as error probabilities are allowed to depend on the disease status d, although estimation of multiple error probabilities could naturally be problematic with limited validation data. Our method is fairly general to accommodate other types of validation designs where infallible data may be available from a 'gold standard' genotyping method, as discussed in [27] and as presented in our real-data example.

3. SIMULATION STUDIES

In this section, we present numerical evidence in the form of simulation studies to illustrate the advantage of our proposed methods. Generally, we assume that the genetic variant of interest is a biallelic locus with the wild- and variant-type alleles and consider a dominant model for the effect of the gene variant. We also assume a binary environmental exposure and consider a commonly prevalent exposure. We follow a similar simulation design as mentioned in [28].

We first generate the parental genotype data for each individual. To accomplish this for each family F, we simulate a family-specific allele frequency parameter $\theta_F = \exp(\mu_F)/\{1+\exp(\mu_F)\}$, where μ_F is generated from a normal distribution with mean μ and variance σ^2. We select μ in such a way that the marginal probability of the genotype variant of interest (assuming a dominant model) is fixed at a given prevalence value in the generated population. We consider two situations with $P(G=1)$ fixed at 0.2 and 0.05, to represent a common and a rare gene, respectively. Given the allele frequency parameter θ_F, we generate the genotype data for the parents, assuming Hardy–Weinberg equilibrium and that the parents are independent. Given the genotypes of the parents, we generate the genotype for one offspring based on a standard Mendelian mode of inheritance. We independently generate the environmental exposure for this offspring based on the marginal probability of exposure ($E=1$) for the underlying population. Given the information on genetic and environmental factors, we generate the disease outcome for each individual, independent of others, using the logistic regression model for disease risk

$$\log\left\{\frac{P(D=1|G,E)}{P(D=0|G,E)}\right\} = \beta_0 + \beta_E * E + \beta_G * G + \beta_{GE} G \times E \tag{13}$$

We choose the main effect parameters $\beta_E = \log(\mathrm{OR}_{10}) = \log(2)$ and $\beta_G = \log(\mathrm{OR}_{01}) = \log(2)$ and consider a multiplicative interaction between G and E, fixing $\beta_{GE} = \log(\psi) = \log(2)$. We select the value of β_0 so that the marginal probability of the disease in the population $P(D=1) \approx 0.01$. Following this scheme, we first generate data for a large number of individuals, which we treat as the underlying population, and then randomly select 1000 cases and 1000 controls from this population.

We then retain the disease, genotype and environmental exposure information and discard the rest of the data. Following the definitions of sensitivity and specificity, we randomly misclassify the genotype and environmental exposure information, independent of one another, but keep the disease information unchanged. We set the specificity of the instruments at 1 and consider the following settings: (1) $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$ (2) $se_{0G} = se_{1G} = 0.9$ and $se_{0E} = se_{1E} = 0.8$.

For each scenario, we simulate 500 data sets and analyze the data by implementing the adjusted formulation, but first assuming the true sensitivity and specificity of the genetic and environmental factors to be known. We compare the results both in the absence and in the presence of misclassification (unadjusted and adjusted). We apply our method under all three models as discussed in

Copyright © 2007 John Wiley & Sons, Ltd.

Table IV. Results for 500 simulated unmatched case–control data sets (750/750), where specificity for both genetic and environmental factors $=1.0$, $se_{0G}=se_{1G}=0.95$ and $se_{0E}=se_{1E}=0.9$.

			OR_{10}	OR_{01}	ψ	Power
Constraints	Misclassification		2.0000	2.0000	2.0000	$H_0:\psi=1$
None	No	MLE	2.0238	2.0344	1.9897	0.7660
		s.e.	0.2753	0.3874	0.4853	
		MSE	0.0770	0.1708	0.2587	
		Coverage	0.9448	0.9518	0.9500	
	Yes and unadjusted	MLE	1.8788	2.2731	1.7084	0.5840
		s.e.	0.2462	0.3992	0.4085	
		MSE	0.0752	0.2462	0.2562	
		Coverage	0.9160	0.8800	0.8620	
	Yes and adjusted	MLE	2.0125	2.0246	2.0430	0.6460
		s.e.	0.3168	0.4485	0.6004	
		MSE	0.0997	0.2127	0.3771	
		Coverage	0.9340	0.9620	0.9540	
G–E independence and rare disease	No	MLE	1.9916	1.9980	1.9848	0.9714
		s.e.	0.2589	0.3385	0.3359	
		MSE	0.0656	0.1178	0.1128	
		Coverage	0.9763	0.9472	0.9499	
	Yes and unadjusted	MLE	1.8787	2.2755	1.6640	0.8997
		s.e.	0.2355	0.3550	0.2614	
		MSE	0.0675	0.2056	0.1803	
		Coverage	0.9208	0.8786	0.7573	
	Yes and adjusted	MLE	2.0109	2.0194	1.9866	0.8892
		s.e.	0.3032	0.3991	0.4270	
		MSE	0.0882	0.1678	0.1806	
		Coverage	0.9604	0.9472	0.9420	
G–E independence and $P(D=1)$ known	No	MLE	2.0020	2.0005	1.9900	0.9728
		s.e.	0.2604	0.3388	0.3402	
		MSE	0.0649	0.1150	0.1175	
		Coverage	0.9736	0.9446	0.9604	
	Yes and unadjusted	MLE	1.8713	2.2567	1.6957	0.9129
		s.e.	0.2345	0.3521	0.2699	
		MSE	0.0690	0.1936	0.1643	
		Coverage	0.9103	0.8918	0.7968	
	Yes and adjusted	MLE	1.9998	1.9951	2.0371	0.9103
		s.e.	0.3016	0.3950	0.4432	
		MSE	0.0872	0.1641	0.1954	
		Coverage	0.9604	0.9446	0.9472	

$P(D=1)\approx 0.01$, $P(E=1)\approx 0.5$ and $P(G=1)\approx 0.2$. s.e. refers to the average standard error of the odds ratio estimates. Mean-squared error (MSE) is estimated based on the average squared deviations of the 500 estimates from their true value.

Section 2.1. Tables IV–VI summarize the results of analyzing unmatched case–control data with different sample sizes (1000/1000 and 750/750) for different choices of misclassification error rates with $P(E=1)=0.5$ and $P(G=1)=0.2$.

To summarize, in the presence of misclassification, the estimates without adjustment show high bias and have significantly large mean-squared errors (MSEs), but the standard errors are not necessarily larger when compared with the estimates in the absence of misclassification. We observe that the estimates of ψ without adjustment are biased towards the null. The adjusted estimates that

Table V. Results of unmatched case–control data (1000/1000), where specificity for both genetic and environmental factors = 1.0, $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$.

			OR_{10}	OR_{01}	ψ	Power
Constraints	Misclassification		2.0000	2.0000	2.0000	$H_0 : \psi = 1$
None	No	MLE	2.0059	2.0295	1.9755	0.8580
		s.e.	0.2361	0.3345	0.4168	
		MSE	0.0499	0.1162	0.1779	
		Coverage	0.9520	0.9580	0.9380	
	Yes and unadjusted	MLE	1.8855	2.2858	1.6823	0.6800
		s.e.	0.2139	0.3473	0.3482	
		MSE	0.0570	0.2076	0.2251	
		Coverage	0.9140	0.8660	0.8200	
	Yes and adjusted	MLE	2.0205	2.0383	2.0001	0.7380
		s.e.	0.2753	0.3901	0.5082	
		MSE	0.0736	0.1654	0.2722	
		Coverage	0.9620	0.9500	0.9540	
G–E independence and rare disease	No	MLE	1.9953	2.0070	1.9329	0.9906
		s.e.	0.2244	0.2937	0.2823	
		MSE	0.0480	0.0959	0.0923	
		Coverage	0.9519	0.9412	0.9225	
	Yes and unadjusted	MLE	1.8797	2.2712	1.6308	0.9492
		s.e.	0.2038	0.3067	0.2215	
		MSE	0.0551	0.1578	0.1841	
		Coverage	0.9144	0.8663	0.6364	
	Yes and adjusted	MLE	2.0141	2.0212	1.9295	0.9492
		s.e.	0.2628	0.3454	0.3573	
		MSE	0.0685	0.1069	0.1307	
		Coverage	0.9519	0.9759	0.9519	
G–E independence and $P(D=1)$ known	No	MLE	1.9863	1.9876	1.9728	0.9938
		s.e.	0.2234	0.2908	0.2913	
		MSE	0.0477	0.0944	0.0943	
		Coverage	0.9520	0.9547	0.9360	
	Yes and unadjusted	MLE	1.8724	2.2527	1.6613	0.9599
		s.e.	0.2030	0.3042	0.2287	
		MSE	0.0566	0.1467	0.1657	
		Coverage	0.9064	0.8797	0.6898	
	Yes and adjusted	MLE	2.0041	1.9995	1.9765	0.9600
		s.e.	0.2616	0.3423	0.3705	
		MSE	0.0677	0.1062	0.1365	
		Coverage	0.9493	0.9707	0.9600	

$P(D=1) \approx 0.01$, $P(E=1) \approx 0.5$ and $P(G=1) \approx 0.2$. s.e. refers to the average standard error of the odds ratio estimates. Mean-squared error (MSE) is estimated based on the average squared deviations of the 500 estimates from their true value.

are obtained through our proposed formulation are quite close to the true parameters, except with relatively large standard errors.

In the absence of misclassification, the models under the independence assumption provide much more precise estimates, i.e. smaller standard errors and MSEs, which is now a well-established observation in the literature [15, 16]. Significant gain in efficiency continues to be maintained by the constrained MLEs when both estimates are corrected for misclassification error. A point worth noting is that, under the independence model, the coverage probability of the confidence interval

Copyright © 2007 John Wiley & Sons, Ltd.

Table VI. Results of unmatched case–control data (750/750), where specificity for both genetic and environmental factors $= 1.0$, $se_{0G} = se_{1G} = 0.9$ and $se_{0E} = se_{1E} = 0.8$.

			OR_{10}	OR_{01}	ψ	Power
Constraints	Misclassification		2.0000	2.0000	2.0000	$H_0 : \psi = 1$
None	No	MLE	2.0209	2.0392	1.9681	0.7820
		s.e.	0.2750	0.3892	0.4798	
		MSE	0.0693	0.1534	0.2149	
		Coverage	0.9600	0.9580	0.9540	
	Yes and unadjusted	MLE	1.7746	2.4391	1.5212	0.3620
		s.e.	0.2278	0.4042	0.3660	
		MSE	0.0981	0.3551	0.3599	
		Coverage	0.8500	0.8000	0.7740	
	Yes and adjusted	MLE	2.0117	2.0508	2.0532	0.4620
		s.e.	0.3669	0.5317	0.7268	
		MSE	0.1210	0.2746	0.5381	
		Coverage	0.9500	0.9620	0.9520	
G–E independence and rare disease	No	MLE	2.0112	2.0211	1.9488	0.9714
		s.e.	0.2617	0.3424	0.3296	
		MSE	0.0656	0.1178	0.1128	
		Coverage	0.9642	0.9427	0.9283	
	Yes and unadjusted	MLE	1.7949	2.4612	1.4584	0.6738
		s.e.	0.2203	0.3631	0.2203	
		MSE	0.0891	0.3460	0.3486	
		Coverage	0.8423	0.7312	0.4516	
	Yes and adjusted	MLE	2.0456	2.0927	1.9369	0.6738
		s.e.	0.3581	0.4797	0.5194	
		MSE	0.1253	0.2646	0.3227	
		Coverage	0.9534	0.9462	0.8961	
G–E independence and $P(D=1)$ known	No	MLE	2.0020	2.0005	1.9900	0.9728
		s.e.	0.2604	0.3388	0.3402	
		MSE	0.0649	0.1150	0.1175	
		Coverage	0.9606	0.9391	0.9391	
	Yes and unadjusted	MLE	1.7890	2.4454	1.4830	0.7025
		s.e.	0.2195	0.3608	0.2275	
		MSE	0.0912	0.3303	0.3265	
		Coverage	0.8387	0.7384	0.4946	
	Yes and adjusted	MLE	2.0328	2.0646	1.9934	0.6738
		s.e.	0.3560	0.4748	0.5418	
		MSE	0.1232	0.2564	0.3492	
		Coverage	0.9498	0.9427	0.9068	

$P(D=1)\approx 0.01$, $P(E=1)\approx 0.5$ and $P(G=1)\approx 0.2$. s.e. refers to the average standard error of the odds ratio estimates. Mean-squared error (MSE) is estimated based on the average squared deviations of the 500 estimates from their true value.

for the interaction parameter is greatly affected if one fails to account for the misclassification error. As we expected, a larger sample size improves the power of testing the interaction effect as well as the precision of the parameter estimates.

To evaluate the performance of the proposed method for a different allele frequency of the genetic marker, we carried out two sets of simulation in Table VII which share an identical simulation scenarios, except one with a more common gene $P(G=1)=0.20$ and the other with a relatively

Table VII. Results of unmatched case–control data (750/750), where specificity for environmental factor $= 1.0$, $sp_{0G} = sp_{1G} = 0.95$, $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$.

$P(G=1)$	Constraints	Misclassification		$\mathrm{OR}_{10}=2$	$\mathrm{OR}_{01}=2$	$\psi=2$	Power of $H_0: \psi=1$
0.2	None	No	MLE	1.9954	1.9739	2.0180	0.8300
			s.e.	0.2709	0.3756	0.4915	
			MSE	0.0659	0.1255	0.1890	
			Coverage	0.9800	0.9700	0.9800	
		Yes	MLE	1.8743	2.0449	1.6533	0.5300
		and unadjusted	s.e.	0.2517	0.3473	0.3806	
			MSE	0.0722	0.1285	0.2521	
			Coverage	0.9200	0.9800	0.8300	
		Yes	MLE	2.0046	1.9993	2.0680	0.5600
		and adjusted	s.e.	0.3249	0.4838	0.6635	
			MSE	0.0945	0.2519	0.4006	
			Coverage	0.9700	0.9800	0.97	
	G–E	No	MLE	2.0122	1.9991	1.9278	0.9600
	independence		s.e.	0.2615	0.3381	0.3255	
	and rare disease		MSE	0.0679	0.1069	0.0969	
			Coverage	0.9500	0.9600	0.9200	
		Yes	MLE	1.8820	2.0519	1.6028	0.8400
		and unadjusted	s.e.	0.2396	0.3114	0.2495	
			MSE	0.0672	0.1076	0.2172	
			Coverage	0.9300	0.9600	0.6900	
		Yes	MLE	2.0141	2.0028	1.9779	0.8000
		and adjusted	s.e.	0.3102	0.4280	0.4534	
			MSE	0.0893	0.2098	0.1979	
			Coverage	0.9600	0.9600	0.9700	
0.05	None	No	MLE	1.9829	2.1877	2.0686	0.3434
			s.e.	0.2224	0.7305	0.8959	
			MSE	0.0518	0.6757	0.8075	
			Coverage	0.9495	0.9596	0.9697	
		Yes	MLE	1.8432	1.7344	1.5358	0.2000
		and unadjusted	s.e.	0.2078	0.4168	0.5091	
			MSE	0.0719	0.2582	0.5218	
			Coverage	0.8400	0.8900	0.8100	
		Yes	MLE	2.0139	2.5259	2.1739	0.1674
		and adjusted	s.e.	0.2647	1.5286	1.7194	
			MSE	0.0802	1.9632	1.8531	
			Coverage	0.9551	0.9655	0.9605	
	G–E	No	MLE	1.9845	2.1302	1.9563	0.7200
	independence		s.e.	0.2202	0.5959	0.5097	
	and rare disease		MSE	0.0517	0.4531	0.3223	
			Coverage	0.9400	0.9600	0.9300	
		Yes	MLE	1.8438	1.7194	1.4724	0.4600
		and unadjusted	s.e.	0.2033	0.3607	0.3047	
			MSE	0.0710	0.2117	0.4003	
			Coverage	0.8500	0.8400	0.6100	
		Yes	MLE	2.0078	2.3675	2.0274	0.3959
		and adjusted	s.e.	0.2579	1.0526	0.8768	
			MSE	0.0771	1.2476	0.6625	
			Coverage	0.9490	0.9694	0.8980	

$P(D=1) \approx 0.01$, $P(E=1) \approx 0.5$, but with different values of $P(G=1)$. s.e. refers to the average standard error of the odds ratio estimates. Mean-squared error (MSE) is estimated based on the average squared deviations of the 500 estimates from their true value.

Copyright © 2007 John Wiley & Sons, Ltd.

rare mutation $P(G=1)=0.05$. We observe that for the rare gene situation both the corrected and uncorrected estimates have larger standard errors and MSEs for OR_{01} and ψ, although the corrected estimates perform better than the uncorrected ones. The power for testing the interaction effect is significantly less for the rare genetic mutation. However, the models under the independence assumption always provide much more precise estimates, i.e. smaller standard errors and MSEs, as well as significantly larger power.

We also present a set of simulation results assuming the misclassification error rates to be unknown and estimated using a hypothetical validation study independent of the main study.

Table VIII. Results of unmatched case–control data (1000/1000), with no misclassification in G, and $sp_{0E}=0.9$, $sp_{1E}=0.98$, $se_{0E}=0.85$ and $se_{1E}=0.8$.

			OR_{10}	OR_{01}	ψ	Power
Constraints	Misclassification error rates		2.0000	2.0000	2.0000	$H_0:\psi=1$
None	True	MLE	2.0167	2.0330	2.0323	0.6300
		s.e.	0.3220	0.4372	0.6031	
		MSE	0.0955	0.1803	0.3160	
		Coverage	0.9580	0.9548	0.9536	
	Wrong guess	MLE	3.1183	1.2462	3.9070	0.7137
		s.e.	0.6395	0.5184	2.5693	
		MSE	1.6278	0.8167	9.5234	
		Coverage	0.4581	0.9759	0.9604	
	Estimated only by validation data	MLE	2.1278	1.9778	2.2087	0.6283
		s.e.	0.3631	0.4528	0.7244	
		MSE	0.8178	0.2555	0.7040	
		Coverage	0.6681	0.9358	0.9202	
	Estimated by using the joint likelihood	MLE	2.1529	1.9386	2.4922	0.6283
		s.e.	0.3783	0.4587	0.7834	
		MSE	0.8755	0.2685	0.8855	
		Coverage	0.6881	0.9513	0.9646	
G–E independence and rare disease	True	MLE	2.0174	2.0182	1.9958	0.8894
		s.e.	0.3088	0.3812	0.4324	
		MSE	0.0920	0.1400	0.1555	
		Coverage	0.9735	0.9690	0.9602	
	Wrong guess	MLE	3.1209	1.2409	3.8632	0.8496
		s.e.	0.6279	0.5027	2.4319	
		MSE	1.6327	0.8208	8.7744	
		Coverage	0.4204	0.9646	0.9867	
	Estimated only by validation data	MLE	2.0908	1.9964	2.0859	0.8898
		s.e.	0.3351	0.3866	0.5028	
		MSE	0.5377	0.2079	0.3764	
		Coverage	0.8660	0.9324	0.9365	
	Estimated by using the joint likelihood	MLE	2.1169	1.9552	2.1586	0.8689
		s.e.	0.3490	0.3921	0.5477	
		MSE	0.5757	0.2226	0.5064	
		Coverage	0.8783	0.9562	0.9530	

$P(D=1)\approx 0.01$, $P(E=1)\approx 0.5$ and $P(G=1)\approx 0.2$. The second column refers to different methods of estimating the misclassification error rates. The randomly guessed values of the error rates were $sp_{0E}=0.95$, $sp_{1E}=0.95$, $se_{0E}=0.9$ and $se_{1E}=0.7$. s.e. refers to the average standard error of the odds ratio estimates. Mean-squared error (MSE) is estimated based on the average squared deviations of the 500 estimates from their true value.

Copyright © 2007 John Wiley & Sons, Ltd.

We assume no misclassification in one of the factors, say G, and $sp_{0E}=0.9$, $sp_{1E}=0.98$, $se_{0E}=0.85$ and $se_{1E}=0.8$. We simulate 1000 data sets and analyze the data by implementing the adjusted formulation and comparing the results by:

(1) plugging in the true known error rates, i.e. $sp_{0E}=0.9$, $sp_{1E}=0.98$, $se_{0E}=0.85$ and $se_{1E}=0.8$;
(2) plugging in a set of randomly guessed error rates, e.g. $sp_{0E}=0.95$, $sp_{1E}=0.95$, $se_{0E}=0.9$ and $se_{1E}=0.7$;
(3) plugging in the error rates estimated by the validation data *only* as discussed in (i) of Section 2.4;
(4) estimating all the parameters based on the full joint likelihood as in option (ii) (d) in Section 2.4.

Table VIII summarizes the results under the unconstrained model and rare disease and G–E independence assumption. We note that when the misclassification error rates are estimated separately from the validation study, the coverage probabilities of the confidence intervals are slightly smaller than the nominal confidence level. This is because we ignore the uncertainty in the estimation of the error rates, but for all practical inferential purposes the plug-in method may be quite acceptable due to its ease of implementation. The point estimates will suffer greatly if one uses wrong guesses for the error probabilities. As noticed in much of the measurement error literature, the bias in the estimates is corrected, but the standard error typically increases due to the correction, and often there is no significant gain in terms of the power of the testing procedure.

4. REAL DATA ANALYSIS

To illustrate the use of these methods in a real setting, we analyze data from a case–control study of colorectal cancer [29]. The aims of the study were to assess the effects of genes, diet and the interaction between both on the risk of colorectal cancer. All cases diagnosed of colorectal cancer in a university hospital in Barcelona during 1996–1998 were included. For each case, a frequency-matched control was selected among the patients of the same hospital (for study details, see [30]). All subjects were interviewed to assess risk factors, including diet, and they provided a blood sample for genetic analysis. For the purpose of this example, we have selected the study of *SULT1A1* (phenol sulfotransferase), a gene highly expressed in the colon that metabolizes drugs, hormones, some nutrients and other xenobiotics. We were interested in the possibility that the risk associated with this gene, if any, could be related to diet. Hence, we explored the interactions with nutrients estimated from a food frequency questionnaire and found that zinc intake could be a potential modifier. Since this was unexpected, and further analysis and studies were needed to confirm this interaction, it is a good example to illustrate the methods. Initially, a polymorphism in *SULT1A1* was genotyped by very precise and well-tested methods in 293 cases and 272 controls [30]. Later, an extended sample (377 cases and 326 controls) was genotyped for a large selection of polymorphisms in metabolism genes, which included *SULT1A1*. This latter analysis used a microarray genotyping method that had been validated and was known to have good although not perfect, accuracy [29]. We considered the first genotype results as gold standard and used them to calibrate the odds ratios for the analysis of interaction between *SULT1A1* (G) and zinc intake (E). Zinc intake was dichotomized by individuals taking less or more than the median value of the zinc intake, this median being determined from the sample data. From a scientific point of

Copyright © 2007 John Wiley & Sons, Ltd.

Table IX. Results of analysis of the real data.

Constraints	Misclassification		OR_{10}	OR_{01}	ψ
None	Unadjusted	MLE	0.4763	0.7334	1.8929
		s.e.	0.1004	0.1616	0.5975
		CI	(0.3151, 0.7200)	(0.4762, 1.1297)	(1.0196, 3.5142)
	Adjusted	MLE	0.4726	0.7592	1.9406
		s.e.	0.1018	0.1749	0.6432
		CI	(0.3099, 0.7208)	(0.4833, 1.1926)	(1.0135, 3.7159)
IR	Unadjusted	MLE	0.6011	0.9746	1.1180
		s.e.	0.1103	0.1780	0.2417
		CI	(0.4195, 0.8613)	(0.6814, 1.3941)	(0.7318, 1.7078)
	Adjusted	MLE	0.5987	1.0177	1.1265
		s.e.	0.1124	0.1960	0.2603
		CI	(0.4144, 0.8649)	(0.6977, 1.4844)	(0.7163, 1.7717)

s.e. refers to the standard error of the odds ratio estimates. Confidence intervals for the odds ratios are calculated by exponentiating the CI based on the asymptotic normality of the MLEs of log OR.

Table X. Results of analysis of the real data with 10 per cent artificial misclassification, introduced at random.

Constraints	Misclassification		OR_{10}	OR_{01}	ψ
None	Unadjusted	MLE	0.5202	0.8708	1.7651
		s.e.	0.1189	0.2120	0.6102
		CI	(0.3324, 0.8141)	(0.5404, 1.4031)	(0.8964, 3.4759)
	Adjusted	MLE	0.4922	0.8040	2.0084
		s.e.	0.1255	0.2459	0.8694
		CI	(0.2986, 0.8113)	(0.4415, 1.4641)	(0.8597, 4.6918)
IR	Unadjusted	MLE	0.5689	0.9781	1.4236
		s.e.	0.1156	0.1990	0.3414
		CI	(0.3820, 0.8474)	(0.6564, 1.4573)	(0.8897, 2.2779)
	Adjusted	MLE	0.5542	0.9369	1.5118
		s.e.	0.1185	0.2349	0.4250
		CI	(0.3644, 0.8428)	(0.5732, 1.5315)	(0.8714, 2.6231)

The point estimates of specificities and sensitivities are $sp_{0G}=0.9$, $se_{0G}=0.857$, $sp_{1G}=0.933$, $se_{1G}=0.923$. Confidence intervals for the odds ratios are calculated by exponentiating the CI based on the asymptotic normality of the MLEs of log OR.

view, it is reasonable to assume that *SULT1A1* mutation status and zinc intake are independent, and colorectal cancer is a rare disease in the study population with an estimated crude annual incidence rate of 46 cases per 100 000 people and a lifetime cumulative risk of 5 per cent in Spain.

In many common situations, the 'gold standard' exposure measurements, namely the true G^* and E^*, may be available only for a subset $\mathbf{V}$ of the original sample $\mathbf{S}$, and we have complete data D, G, G^*, E^*, for that sub-sample and reduced data D, G, E on the remaining sample $\mathbf{S}-\mathbf{V}$. In that case, the joint likelihood is of a that of different form that of joint likelihood given in (10)

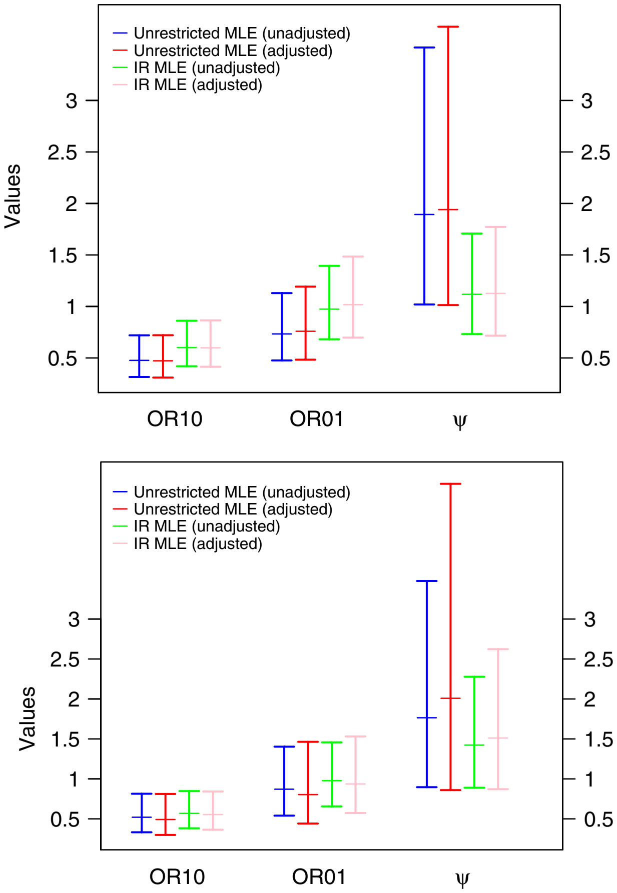

Figure 1. Results of analysis of colorectal cancer study: OR_{10}, OR_{01} and ψ represent the main effect of zinc intake, the main effect of *SULT1A1* and their interaction, respectively. Confidence intervals for the odds ratios are calculated by exponentiating the CI based on the asymptotic normality of the MLEs of $\log \mathrm{OR}$. The top figure is based on real data, whereas the bottom figure is based on real data, with additional 10 per cent misclassification of genotypes introduced at random.

and is given by

$$L_{\text{full}} = \prod_{i \in \mathbf{V}} P(G_i, G_i^*, E_i, E_i^* | D_i) \times \prod_{i \in \mathbf{S}-\mathbf{V}} P(G_i, E_i | D_i) \tag{14}$$

The second term involving reduced data is expressed exactly in the same way as we factorize L_2 in Result 1, summing over the latent true measurements G^* and E^*. The first term with complete data is factorized similarly, simply omitting the sum over the latent values of G^* and E^*, which is not needed for complete data as we have perfect true measurements for this subsample. We then maximize the joint likelihood in terms of all model parameters and error rates.

In our real data example, we have G^* available for a subsample and ignore the possible misclassification of E in the absence of any validation data on E. The crude estimates of agreement between

G and G^* obtained only from the validation data are $sp_{0G} = 0.981$, $sp_{1G} = 0.975$, $se_{0G} = 0.991$ and $se_{1G} = 0.959$. The results from the plug-in method are presented in Table AI. One can notice the dramatic change in the width of the confidence interval for the interaction parameter, when independence assumption is exploited. The inference results from the IR assumption indicate that the effect of zinc intake is protective, whereas carrier status for *SULT1A1* as well as the interaction parameter are not statistically significant. Note that, using the unconstrained ML method, the interaction odds ratio is detected to be marginally significant (CI (1.02, 3.51)), which is most likely to be a false positive from an epidemiological perspective. There are certain numerical differences between the estimates adjusted for misclassification and unadjusted estimates in the real data, but the inference is the same. To explore further the role of misclassification, we randomly selected 10 per cent of the observations where G and G^* agreed and changed the value of G to create artificial misclassification. The results are presented in Tables IX and X. The difference in the inference shows that under misclassification, the odds ratio estimates are indeed biased towards the null, and our simple method corrects for such attenuation effects. Figure 1 presents the OR estimates and the confidence intervals for the two misclassification scenarios.

5. CONCLUSION

We describe a relatively simple analytical formulation to account for misclassification of exposures in studies of gene–environment interaction based on sensitivity and specificity of the measurement instrument for genetic and environmental factors in unmatched case–control studies. As illustrated in our simulations, even with relatively small degrees of error (i.e. sensitivity or specificity quite close to 1), the estimates of the parameters of interest could have relatively large biases. Our corrected estimates minimize the biases and are found to be closer to the true parameters, with better MSE properties than their unadjusted counterparts.

This paper presents a clear insight into how ML estimation under a constrained exposure space leads to efficiency advantages, even in the absence of a perfectly measured data set. The 2×4 table is a pivotal and routinely used tool in studies of gene–environment interaction [21], and we present a comprehensive treatment of this case with closed-form expressions for the estimates and the standard errors where available. The proposed method can be easily extended to genotype data which inherently have three levels, that is, when we have a 2×6 table instead of a 2×4 table (Appendix A.3). According to the results of the simulation, one gains significant efficiency under the G–E independence assumption and after adjusting for misclassification errors. However, cautions regarding the validity of the independence assumption should be exercised while using the proposed methods [16, 31]. Extending the methods under the conditional independence of G and E as in [19], but not just restricted to a case-only design as in [19], could be one avenue to dealing with violation of the independence assumption.

To conclude, we would like to mention that it is often hard to obtain a precise estimate of the interaction odds ratio due to sparsity of observations in some rare genotype–exposure configurations, and the use of G–E independence assumption and adjusting for misclassification errors give us a better chance to detect G–E interaction. Studying the synergism of gene and the environment in the etiology of rare and complex diseases is an important problem in modern genetic epidemiology, and efficient estimation strategies are much needed. This paper introduces certain simple and practically useful ideas in the context of this problem, which account for potential genotyping and exposure misclassification errors. The simplicity of the method makes it

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

a viable screening tool for gene–environment interactions in large-scale genomewide association studies.

APPENDIX

Note: All parameters are defined the same as in the text, except those defined separately here. Let $P(G=1)=q_G$ and $P(E=1)=q_E$.

A.1. The constrained ML equations under G–E independence and rare disease assumptions are obtained by differentiating the logarithm of the likelihood (1) with respect to the corresponding parameters:

$$p^2(p_{04}r_{01}-p_{01}r_{04})+pp_{01}p_{04}(r_{11}-\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi r_{14})=p_{01}p_{04}(1-\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi)q$$

$$p^2(p_{04}r_{02}-p_{02}r_{04})+pp_{02}p_{04}\,\mathrm{OR}_{10}(r_{12}-\mathrm{OR}_{01}\,\psi r_{14})=p_{02}p_{04}(\mathrm{OR}_{10}-\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi)q$$

$$p^2(p_{04}r_{03}-p_{03}r_{04})+pp_{03}p_{04}\,\mathrm{OR}_{01}(r_{13}-\mathrm{OR}_{10}\,\psi r_{14})=p_{03}p_{04}(\mathrm{OR}_{01}-\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi)q$$

$$\frac{r_{12}+r_{14}}{\mathrm{OR}_{10}}-\frac{(p_{02}+p_{04}\,\mathrm{OR}_{01}\,\psi)n_1}{p}=0 \tag{A1}$$

$$\frac{r_{13}+r_{14}}{\mathrm{OR}_{01}}-\frac{(p_{03}+p_{04}\,\mathrm{OR}_{10}\,\psi)n_1}{p}=0 \tag{A2}$$

$$\frac{r_{14}}{\psi}-\frac{p_{04}\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,n_1}{p}=0 \tag{A3}$$

where $q=p_{01}r_{11}+p_{02}\,\mathrm{OR}_{10}\,r_{12}+p_{03}\,\mathrm{OR}_{01}\,r_{13}+p_{04}\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi r_{14}$. Recall $p_{04}=1-p_{01}-p_{02}-p_{03}$ and $p=p_{01}+p_{02}\,\mathrm{OR}_{10}+p_{03}\,\mathrm{OR}_{01}+p_{04}\,\mathrm{OR}_{10}\,\mathrm{OR}_{01}\,\psi$. The solutions to the above equations are subjected to the restriction of $p_{01}p_{04}=p_{02}p_{03}$. In the following, we show how to obtain those restricted MLEs:

(1) Plugging (A1)–(A3), we

$$\begin{aligned}\hat{p}&=\hat{p}_{01}+\widehat{\mathrm{OR}}_{10}(\hat{p}_{02}+\hat{p}_{04}\,\widehat{\mathrm{OR}}_{01}\,\hat{\psi})+\widehat{\mathrm{OR}}_{01}(\hat{p}_{03}+\hat{p}_{04}\,\widehat{\mathrm{OR}}_{10}\,\hat{\psi})-\hat{p}_{04}\,\widehat{\mathrm{OR}}_{10}\,\widehat{\mathrm{OR}}_{01}\,\hat{\psi}\\&=\hat{p}_{01}+\frac{r_{12}+r_{14}}{n_1}\hat{p}+\frac{r_{13}+r_{14}}{n_1}\hat{p}-\frac{r_{14}}{n_1}\hat{p}\\&=\hat{p}_{01}+\frac{n_1-r_{11}}{n_1}\hat{p}\end{aligned}$$

Thus,

$$\frac{r_{11}}{n_1}\hat{p}=\hat{p}_{01}\quad\text{and}\quad\hat{p}_{11}=\frac{\hat{p}_{01}}{\hat{p}}=\frac{r_{11}}{n_1}$$

Also, by (A1)–(A3), we can obtain $\hat{p}_{1j}=r_{1j}/n_1$, $j=2,3,4$.

(2) Thus, we can write the profile likelihood as

$$L_p(\mathbf{p}_0,\hat{\mathbf{p}}_1)=\prod_{j=1}^{4}{p_{0j}}^{r_{0j}}\prod_{j=1}^{4}\hat{p}_{1j}^{r_{1j}}\propto\prod_{j=1}^{4}{p_{0j}}^{r_{0j}}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

By the G–E independence and rare disease assumptions, we have

$$\begin{aligned} p_{01} &= (p_{01}+p_{02})(p_{01}+p_{03}) \\ p_{02} &= (p_{01}+p_{02})(p_{02}+p_{04}) \\ p_{03} &= (p_{01}+p_{03})(p_{03}+p_{04}) \end{aligned} \tag{A4}$$

Hence, writing $L_p(\mathbf{p}_0, \hat{\mathbf{p}}_1) \propto (p_{01}+p_{02})^{r_{01}+r_{02}}(p_{01}+p_{03})^{r_{01}+r_{03}}(p_{02}+p_{04})^{r_{02}+r_{04}}(p_{03}+p_{04})^{r_{03}+r_{04}}$, we have

$$\widehat{p_{01}^{IR}+p_{02}^{IR}} = \frac{r_{01}+r_{02}}{n_0}$$

$$\widehat{p_{01}^{IR}+p_{03}^{IR}} = \frac{r_{01}+r_{03}}{n_0}$$

$$\widehat{p_{02}^{IR}+p_{04}^{IR}} = \frac{r_{02}+r_{04}}{n_0}$$

$$\widehat{p_{03}^{IR}+p_{04}^{IR}} = \frac{r_{03}+r_{04}}{n_0}$$

Plugging into (A4), we have (5).

The estimated asymptotic variance–covariance matrix can be obtained by the inverse of the observed information matrix. The observed information matrix is constructed by taking the second derivative of the log-likelihood with respect to the parameters and evaluating them at the MLEs of the parameters, which are the solutions to the above equations.

Here we state how we use the delta method along with the properties of a multinomial distribution and a binomial distribution to obtain the estimated asymptotic variance of the odds ratios.

First we consider OR_{10}, whose MLE is

$$\widehat{\mathrm{OR}_{10}^{IR}} = \frac{r_{12}}{r_{11}} \cdot \frac{r_{01}+r_{03}}{r_{02}+r_{04}} = \frac{\hat{p}_{12}}{\hat{p}_{11}} \cdot \frac{\hat{p}_{01}+\hat{p}_{03}}{\hat{p}_{02}+\hat{p}_{04}}$$

where $\hat{p}_{dj} = r_{dj}/n_d$ ($d=0,1$ and $j=1,2,3,4$). Let $\phi = \hat{p}_{01}+\hat{p}_{03}$; we artificially build a distribution P_A which includes independent distributions P_m and P_b to satisfy this particular odds ratio estimate, $P_A = P_m P_b \propto p_{11}^{r_{11}} p_{12}^{r_{12}} (1-p_{11}-p_{12})^{r_{13}+r_{14}} \phi^{r_{01}+r_{03}} (1-\phi)^{r_{02}+r_{04}}$. Note that, for the multinomial distribution $P_m \propto p_{11}^{r_{11}} p_{12}^{r_{12}} (1-p_{11}-p_{12})^{r_{13}+r_{14}}$,

$$\begin{pmatrix} \hat{p}_{12} \\ \hat{p}_{11} \end{pmatrix} \sim \mathrm{AN}\left(\begin{pmatrix} p_{12} \\ p_{11} \end{pmatrix}, \Sigma \right) \quad \text{with } \Sigma = \frac{1}{n_1} \begin{pmatrix} p_{12}(1-p_{12}) & -p_{11}p_{12} \\ -p_{11}p_{12} & p_{11}(1-p_{11}) \end{pmatrix}$$

Let $g(x,y) = \log(x) - \log(y)$, then

$$\frac{\partial g(x,y)}{\partial x} = \frac{1}{x} \quad \text{and} \quad \frac{\partial g(x,y)}{\partial y} = -\frac{1}{y}$$

Thus, by the delta method,

$$\log\left(\frac{\hat{p}_{12}}{\hat{p}_{11}}\right) \sim \mathrm{AN}\left(\log\left(\frac{p_{12}}{p_{11}}\right), \left(\frac{1}{p_{12}}, -\frac{1}{p_{11}}\right) \Sigma \left(\frac{1}{p_{12}}, -\frac{1}{p_{11}}\right)^{\mathrm{T}} \right)$$

Copyright © 2007 John Wiley & Sons, Ltd.

Hence, the estimated asymptotic variance of $\log(\hat{p}_{12}/\hat{p}_{11})$ is

$$\widehat{\text{AVAR}}\left(\log\left(\frac{\hat{p}_{12}}{\hat{p}_{11}}\right)\right)=\frac{1}{n_1}\left(\frac{1}{\hat{p}_{11}}+\frac{1}{\hat{p}_{12}}\right)=\frac{1}{r_{11}}+\frac{1}{r_{12}}$$

Similarly, for the binomial distribution $P_b \propto \phi^{r_{01}+r_{03}}(1-\phi)^{r_{02}+r_{04}}$,

$$\frac{\hat{\phi}}{1-\hat{\phi}}\sim \text{AN}\left(\frac{\phi}{1-\phi},\frac{1}{n_0}\left(\frac{1}{\phi^2(1-\phi)^2}\right)\right)$$

Let $g(x)=\log(x)-\log(1-x)$, then

$$\frac{\mathrm{d}g(x)}{\mathrm{d}x}=\frac{1}{x}+\frac{1}{1-x}=\frac{1}{x(1-x)}$$

Thus, by the delta method,

$$\log\left(\frac{\hat{\phi}}{1-\hat{\phi}}\right)\sim \text{AN}\left(\log\left(\frac{\phi}{1-\phi}\right),\frac{1}{n_0}\left(\frac{1}{\phi}+\frac{1}{1-\phi}\right)\right)$$

Hence, the estimated asymptotic variance of $\log(\hat{\phi}/1-\hat{\phi})$ is

$$\widehat{\text{AVAR}}\left(\log\left(\frac{\hat{\phi}}{1-\hat{\phi}}\right)\right)=\frac{1}{n_0}\left(\frac{1}{\hat{\phi}}+\frac{1}{1-\hat{\phi}}\right)=\frac{1}{r_{01}+r_{03}}+\frac{1}{r_{02}+r_{04}}$$

Since $\log(\widehat{\text{OR}}_{10}^{IR})=\log(\hat{p}_{12})-\log(\hat{p}_{11})+\log(\hat{\phi})-\log(1-\hat{\phi})$,

$$\begin{aligned}\widehat{\text{AVAR}}(\log(\widehat{\text{OR}}_{10}^{IR}))&=\widehat{\text{AVAR}}\left(\log\left(\frac{\hat{p}_{12}}{\hat{p}_{11}}\right)\right)+\widehat{\text{AVAR}}\left(\log\left(\frac{\hat{\phi}}{1-\hat{\phi}}\right)\right)\\&=\frac{1}{r_{11}}+\frac{1}{r_{12}}+\frac{1}{r_{01}+r_{03}}+\frac{1}{r_{02}+r_{04}}\end{aligned}$$

Calculation of the estimated asymptotic variances of $\widehat{\text{OR}}_{01}^{IR}$ and $\hat{\psi}^{IR}$ is based on the same ideas.

A.2. Obtain restriction (4). Following (3), we have

$$(1-q_G)(1-q_E)=(1-\pi)p_{01}+\pi p_{01}/p \tag{A5}$$

$$(1-q_G)q_E=(1-\pi)p_{02}+\text{OR}_{10}\,\pi p_{02}/p \tag{A6}$$

$$q_G(1-q_E)=(1-\pi)p_{03}+\text{OR}_{01}\,\pi p_{03}/p \tag{A7}$$

$$q_G q_E=(1-\pi)p_{04}+\pi\,\text{OR}_{10}\,\text{OR}_{01}\,\psi p_{04}/p \tag{A8}$$

Note that using (A6)–(A8) we obtain the following two equations:

$$\begin{aligned}q_E&=(1-\pi)p_{02}+\text{OR}_{10}\,\pi p_{02}/p+(1-\pi)p_{04}+\pi\,\text{OR}_{10}\,\text{OR}_{01}\,\psi p_{04}/p\\q_G&=(1-\pi)p_{03}+\text{OR}_{01}\,\pi p_{03}/p+(1-\pi)p_{04}+\pi\,\text{OR}_{10}\,\text{OR}_{01}\,\psi p_{04}/p\end{aligned} \tag{A9}$$

Thus, using (A8) and (A9) we have (4).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table AI. Data for an unmatched case–control study with a three-level genetic factor and a binary environmental exposure.

	$G=0$		$G=1$		$G=2$		
	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	Total
j	1	2	3	4	5	6	
$D=0$	r_{01}	r_{02}	r_{03}	r_{04}	r_{05}	r_{06}	n_0
$D=1$	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	n_1

A.3. Extension to 2×6 tables. Here we briefly describe how to extend our approach to the case of a 2×6 table when the genotype data are recorded as three levels, i.e. 0, 1 and 2, depending on the number of copies of the high-risk allele at a biallelic locus. The data can be represented as in Table AI. In this case, in addition to the parameters OR_{10}, OR_{01} and ψ introduced in Section 2, we introduce one more odds ratio parameter related to the main effect of $G=2$, namely $\mathrm{OR}_{02}=p_{01}p_{15}/(p_{11}p_{05})$ and one more multiplicative interaction parameter $\gamma=\mathrm{OR}_{12}/(\mathrm{OR}_{10}\,\mathrm{OR}_{02})=p_{02}p_{05}p_{11}p_{16}/(p_{01}p_{06}p_{12}p_{15})$ (where $\mathrm{OR}_{12}=p_{01}p_{16}/(p_{11}p_{06})$). The case probabilities can be parameterized in terms of the five control probabilities p_{0j}, $j=1,\ldots,5$, and OR_{10}, OR_{01}, OR_{02}, ψ and γ in a manner exactly similar to that in Section 2.1. We can write the likelihood, as $L(\mathrm{OR}_{10},\mathrm{OR}_{01},\mathrm{OR}_{02},\psi,\gamma,p_{01},p_{02},p_{03},p_{04},p_{05}|\mathbf{r}_0,\mathbf{r}_1)=\prod_{d=0}^{1}\prod_{j=1}^{6}p_{dj}^{r_{dj}}$. Without any model assumptions, we can obtain the MLEs of the parameters of interest by simply maximizing the above likelihood with the constraints being $\sum_{j=1}^{6}p_{0j}=1$ and $p_{0j}>0$ $(j=1,\ldots,6)$. Under G–E independence and rare disease assumption, we would have additional constraints on p_{0j} $(j=1,\ldots,6)$ given by $p_{01}/p_{02}=p_{03}/p_{04}=p_{05}/p_{06}$. The restricted MLEs can be obtained with closed-form expressions that resemble expressions in Table II.

In the presence of misclassification, now we can have all possible misclassifications of genotypes (0 labeled as 1 or 2, 1 labeled as 0 or 2, 2 labeled as 0 or 1). Accordingly, define $\tau_d(g,g^*)=P(G=g|G^*=g^*,D=d)$, $g,g^*=0,1,2$. Note that $\tau_d(0,g^*)=1-\tau_d(1,g^*)-\tau_d(2,g^*)$. Hence, as in Result 1, we can again write the p_{dj}'s in terms of a linear function of the true parameters p^*_{dj} (denoted by $p_{dj}(\mathbf{p}^*_d)$) as defined by the following equation:

$$\begin{pmatrix} p_{d1} & p_{d2}\\ p_{d3} & p_{d4}\\ p_{d5} & p_{d6}\end{pmatrix}=\begin{pmatrix}\tau_d(0,0) & \tau_d(0,1) & \tau_d(0,2)\\ \tau_d(1,0) & \tau_d(1,1) & \tau_d(1,2)\\ \tau_d(2,0) & \tau_d(2,1) & \tau_d(2,2)\end{pmatrix}\begin{pmatrix}p^*_{d1} & p^*_{d2}\\ p^*_{d3} & p^*_{d4}\\ p^*_{d5} & p^*_{d6}\end{pmatrix}\begin{pmatrix} sp_{dE} & 1-sp_{dE}\\ 1-se_{dE} & se_{dE}\end{pmatrix} \tag{A10}$$

The multinomial likelihood in terms of the underlying true parameters is

$$L(\mathrm{OR}^*_{10},\mathrm{OR}^*_{01},\mathrm{OR}^*_{02},\psi^*,\gamma^*,p^*_{01},p^*_{02},p^*_{03},p^*_{04},p^*_{05}|\mathbf{r}_0,\mathbf{r}_1)=\prod_{d=0}^{1}\prod_{j=1}^{6}p_{dj}(\mathbf{p}^*_d)^{r_{dj}}$$

Now, we can obtain the MLEs of the true parameters by maximizing the above likelihood as done in Section 2.2. For estimating the six error probabilities in each disease subgroup with limited validation data, it may be sensible to assume some structures for the error rates in order to reduce the number of parameters. To this end, one may assume the errors to be the same among cases and

Copyright © 2007 John Wiley & Sons, Ltd.

controls and consider an allele-based error model [32] or a symmetric allele dropout error model [33] depending upon the specific genetic application.

ACKNOWLEDGEMENTS

We thank the associate editor and two referees for their insightful comments. The research of Bhramar Mukherjee was partially supported by NSA young investigator grant H98230-06-1-0033. Victor Moreno received funds from the Spanish Secretara de Estado de Universidades e Investigacin, Ministerio de Educacin y Ciencia to pursue this research.

REFERENCES

1. Greenland S. Statistical uncertainty due to misclassification: implications for validation sub-studies. *Journal of Clinical Epidemiology* 1988; **41**:1167–1176.
2. Greenland S. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* 1996; **25**:1107–1115.
3. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/validation study designs. *Journal of the American Statistical Association* 2000; **95**:51–61.
4. Bashir SA, Duffy SW. The correction of risk estimates for measurement error. *Annals of Epidemiology* 1997; **7**:156–164.
5. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*: *Impacts and Bayesian Adjustments.* Chapman & Hall/CRC Press: London, Boca Raton, 2004.
6. Rice K, Holmans P. Allowing for genotyping error in analysis of unmatched case–control studies. *Annals of Human Genetics* 2003; **67**:165–174.
7. Rice K. Full-likelihood approaches to misclassification of a binary exposure in matched case–control studies. *Statistics in Medicine* 2003; **22**:3177–3194.
8. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology* 1983; **12**:93–97.
9. Greenland S. On correcting for misclassification in twin studies and other matched pair studies. *Statistics in Medicine* 1989; **8**:825–829.
10. Armstrong BG, Whittemore AS, Howe GR. Analysis of case–control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine* 1989; **8**:1151–1163.
11. Satten GA, Carroll RJ. Conditional and unconditional categorical regression models with missing covariates. *Biometrics* 2000; **56**:384–388.
12. Paik MC, Sacco R. Matched case–control data analyses with missing covariates. *Applied Statistics* 2000; **49**: 145–156.
13. Huberman M, Langholz B. Application of the missing-indicator method in matched case–control studies with incomplete data. *American Journal of Epidemiology* 1999; **150**:1340–1345.
14. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case–control studies. *Statistics in Medicine* 1994; **13**:153–162.
15. Umbach DM, Weinberg CR. Designing and analyzing case–control studies to exploit independence of genotype and exposure. *Statistics in Medicine* 1997; **16**:1731–1743.
16. Chatterjee N, Carroll R. Semiparametric maximum likelihood estimation exploiting gene–environment independence in case–control studies. *Biometrika* 2005; **92**:399–418.
17. Garcia-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene–environment interactions in case–control studies. *American Journal of Epidemiology* 1998; **147**:426–433.
18. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case–control studies of gene–environment interactions: assessment of bias and sample size. *Cancer Epidemiology*, *Biomarkers and Prevention* 1999; **8**:1043–1050.
19. Cheng KF. Analysis of case-only studies accounting for genotyping error. *The Annals of Human Genetics* 2007; **71**:238–249.
20. Cheng KF. A maximum likelihood method for studying gene–environment interactions under conditional independence of genotype and exposure. *Statistics in Medicine* 2006; **25**:3093–3109.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

21. Botto LD, Khury MJ. Commentary: facing the challenge of gene–environment interaction: the two-by-four table and beyond. *American Journal of Epidemiology* 2001; **153**(10):1016–1020.
22. Gustafson P, Le ND, Saskin R. Case–control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* 2001; **57**:598–609.
23. Cheng KF, Lin WJ. Retrospective analysis of case–control studies when the population is in Hardy–Weinberg equilibrium. *Statistics in Medicine* 2005; **24**:3289–3310.
24. Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association* 2006; **101**:89–102.
25. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358**:1356–1360.
26. Lai R, Zhang H, Yang Y. Repeated measurement sampling in genetic association analysis with genotyping errors. *Genetic Epidemiology* 2007; **31**:143–153.
27. Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V. Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1):Article 26.
28. Chatterjee N, Kalaylioglu Z, Carroll R. Exploiting gene–environment independence in family-based case–control studies: increased power for detecting associations, interactions and joint effects. *Genetic Epidemiology* 2005; **28**:138–156.
29. Landi S, Gemignani F, Moreno V, Gioia-Patricola L, Chabrier A, Guino E, Navarro M, de Oca J, Capella G, Canzian F. A comprehensive analysis of phase I and phase II metabolism gene polymorphisms and risk of colorectal cancer. *Pharmacogenetics and Genomics* 2005; **15**(8):535–546.
30. Moreno V, Glatt H, Guino E, Fisher E, Meinl W, Navarro M, Badosa JM, Boeing H. Bellvitge Colorectal Cancer Study Group. Polymorphisms in sulfotransferases SULT1A1 and SULT1A2 are not related to colorectal cancer. *International Journal of Cancer* 2005; **113**(4):683–686.
31. Mukherjee B, Zhang L, Ghosh M, Sinha S. Bayesian semiparametric analysis of case–control data under conditional gene–environment independence. *Biometrics* 2007; in press. Available at http://www.blackwell-synergy.com/doi/abs/10.1111/j.1541-0420.2007.00750.x.
32. Gordon D, Heath SC, Liu X, Ott J. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics* 2001; **69**:371–380.
33. Morris RW, Kaplan NL. Testing for association with a case-parents design in the presence of genotyping errors. *Genetic Epidemiology* 2004; **26**(2):142–154.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim