The Statistician (1998)
47, *Part* 2, *pp.* 333–347

Combining exposure information from various sources in an analysis of a case–control study

Trivellore E. Raghunathan†

University of Michigan, Ann Arbor, USA

and David S. Siscovick

University of Washington, Seattle, USA

[Received November 1995. Revised November 1997]

Summary. This paper describes a method for estimating disease–exposure odds ratios in a case–control study where information on the exposure variable is available from several, possibly imperfect, sources. A hybrid approach is developed where a Bayesian perspective is used in combining information from multiple sources, although the ultimate analysis of the disease–exposure association is likelihood based and incorporates the design considerations from a frequentist perspective, namely matching cases and controls on the basis of certain characteristics. The basic analytical strategy involves using Gibbs sampling to draw several sets of actual exposure variables at random from their posterior distribution, conditional on the exposure ascertainment from several sources and other pertinent variables. Each set of drawn values of the actual exposure variable and the confounding variables are used as independent variables in a conditional logistic regression model with case–control status as the dependent variable. The resulting point estimates and their covariance matrices are then combined. This method is applied to a population-based case–control study of the risk of primary cardiac arrest and the intake of n-3 polyunsaturated fatty acids derived mainly from fish and seafood, which motivated this research. This hybrid strategy was developed for pragmatic reasons as these data will be used for several analyses from differing perspectives by different analysts. Hence, this paper also reports an evaluation from a frequentist perspective that investigates the sampling properties of estimates so derived through a simulation study that is similar in many respects to the actual data set analysed. These results show that the estimate of the log-odds ratio obtained by using the method described in this paper is better in terms of bias, the mean-square error and the confidence coverage when compared with the estimate obtained by using only one of the several sources as the exposure variable.

Keywords: Bayesian inference; Gibbs sampling; Measurement error model; Odds ratio; Surrogate variables

1. Introduction

In an epidemiological study of the association between the occurrence of a disease and an exposure variable, multiple measurements may be available to ascertain the exposure of interest. In this paper we address the issue of combining information from these sources to estimate the adjusted odds ratio, a parameter that is commonly used to measure disease–exposure association. A particular problem that motivated this research was a population-based case–control study to determine whether dietary intake of the n-3 polyunsaturated fatty acids derived mainly from fish

†*Address for correspondence*: Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA.
E-mail: teraghu@umich.edu

 0039–0526/98/47333

and seafood, eicosapentaenoic acid (EPA) and docosohexaenoic acid (DHA), reduces the risk of primary cardiac arrest (PCA). The details of this study are given in Siscovick *et al*. (1995). In this study, three possible sources were available to assess the primary exposure variable, the dietary intake of n-3 polyunsaturated fatty acids:

(a) a quantitative food frequency questionnaire, the seafood intake scale, administered to control subjects and surviving case subjects,
(b) a similar questionnaire administered independently to each subject's spouse as a surrogate respondent for the subject (i.e. asking about the subject's intake) and
(c) a measure of n-3 polyunsaturated fatty acids in the red cell membrane as a percentage of total fatty acids.

All three measures reflect dietary intake. The first two approaches give the direct measures of exposure in grams of dietary intake of EPA and DHA in the prior month. Furthermore, since these two fatty acids are synthesized at low rates, the cell membrane levels of these fatty acids reflect primarily dietary intake. Though all three measures are related to the actual dietary intake, none of these measures may be taken as a substitute for the actual dietary intake. The subject and spouse responses may have recall and response biases. The red cell membrane fatty acid in contrast is an indirect measure of dietary intake and is subject to individual metabolic effects. Thus, the objective was to address whether these three sources can be combined to obtain a better estimate of the exposure, the *actual* dietary intake and hence also a better estimate of the relative risk (or the adjusted odds ratio) of PCA associated with the actual dietary seafood intake.

There is a vast literature on inference based on imperfect measures using measurement error models. Various methods have been developed using parametric, semiparametric and nonparametric approaches. Fuller (1987) and Carroll *et al*. (1995) have provided a comprehensive review of the literature on several of these approaches. In particular, the logistic model with measurement error in a covariate was considered by Stefanski and Carroll (1989) who used the sufficient statistics to develop consistent estimates when the conditional distribution of the mismeasured covariate X given the actual value T is normal. Under a similar set-up Whittemore (1989) suggested substituting the James–Stein estimate of $E(T|X)$ in the logistic model. Rosner *et al*. (1992) and Carroll and Stefanski (1990) suggested other approaches for estimating $E(T|X)$. Carroll and Wand (1991), Ii (1992) and Roeder *et al*. (1996) developed semiparametric approaches and Pepe and Fleming (1991) developed a nonparametric approach based on estimated likelihoods. Armstrong *et al*. (1989) and Buonaccorsi (1990) also discussed fitting logistic regression models with covariates measured with error by using the discriminant analysis approach (Efron, 1975).

We adopt a parametric approach that is similar in spirit to the methods described in these references but under a more general set-up dictated partly by the type of data being analysed. We also adopt a hybrid approach in developing the estimate of the adjusted odds ratio. A Bayesian model is used to combine information from multiple sources of exposure variables to estimate the actual dietary intake; then this estimate is used in a logistic regression model to estimate the disease–exposure odds ratio by using a likelihood-based method. The hybrid approach proposed distinguishes the two tasks or stages explicitly. The first is that of estimating the actual exposure variable by combining the imperfect sources of information using a set of model assumptions. The next is to use the estimated exposure variable in the usual analysis that would have been carried out if it had been possible to obtain the actual exposure measures but adjusting the standard errors to reflect the estimation in the first stage. The hybrid approach proposed provides some flexibility in developing models that are finely tuned to two stages and in accommodating varied perspectives in the analysis of data from the basic case–control study. For instance, in the

example discussed in this paper, the cases and controls were matched on gender and age (within 7 years), and hence, for those tuned to the likelihood perspective, the preference was to use a conditional likelihood (Breslow and Day, 1980) approach to estimate the adjusted odds ratio and at the same time a desire for a framework to combine information from multiple sources and the associated uncertainties. The hybrid approach that is described in this paper develops and evaluates a procedure to accommodate the differing perspectives.

As we indicate later, it is possible to develop methods that are either fully likelihood based or fully Bayesian under certain conditions. The fully Bayesian analysis, using for example Gibbs sampling (Richardson and Gilks, 1993), has the advantage that all the uncertainties due to estimating the actual exposure variable are incorporated quite easily. The hybrid approach has some advantages from a practical point of view. The explicit modelling of two phases or stages allows us to use different sets of covariates in the model-building process. For instance, the number of years that the subject and his or her spouse were married or whether or not the subject works away from home has a bearing on the measurement error process but possibly none on the disease–exposure relationship. Though an equivalent Bayesian method can be developed that allows this flexibility, the hybrid approach at least has pragmatic appeal. It is possible to develop such methods for the likelihood-based analysis also. The fully likelihood approach may be computationally convenient only if all the imperfect measures are completely observed (i.e. there are no missing values). However, in the data set certain values on one or more imperfect measures were not observed. For instance, the case subjects who died had only two or one measures depending on whether or not blood was drawn from the subject on his or her death.

The hybrid approach can also be viewed as a multiple-imputation analysis with missing data (Rubin, 1987; Raghunathan and Siscovick, 1996). The actual dietary intake of n-3 fatty acids may be viewed as missing on all the individuals and several sets are being imputed conditionally on a certain set of variables measured or observed on those individuals, assuming that the data are missing at random. The imputed values are then used to form completed data sets and the appropriate multiple-imputation analyses of substantive interest are then performed. Such a view is helpful, if the same data set will be used by many analysts looking at various aspects of disease–risk factor relationships, where the actual exposure variables may be used as a confounder or as a primary exposure variable of interest.

The hybrid approach described in this paper needs to be evaluated for pragmatic reasons. For this, we conducted a simulation study to investigate the properties of the estimates when applied repeatedly under similar settings. We generated several data sets under certain model assumptions and applied our procedure to obtain point and interval estimates of the primary parameter of interest. We evaluated the bias and mean-square error of the point estimates and the actual confidence coverage of the interval estimates across the data sets that were simulated.

The rest of the paper is organized into six sections. In Section 2 we briefly describe the case–control study. Section 3 describes the model assumptions for each stage of the analysis. In Section 4 we describe the basics of the estimation method and include a discussion of the fully Bayesian and likelihood-based approaches. Section 5 describes the results from applying the method to the particular case–control example. In Section 6 we describe the results from the simulation study and, finally, Section 7 concludes with a discussion.

2. Description of case–control study

In the population-based case–control study mentioned earlier, the cases were all incident out-of-hospital PCAs attended by paramedics, satisfying the eligibility criteria listed below, that occurred between 1988 and 1994 in King County, the largest county in the state of Washington. The

cases were identified through a review of incident reports filled out by paramedics from Seattle and King County (Washington). In addition to the incident reports, death certificates, medical examiner reports and autopsy reports (when available) were reviewed to confirm the absence of evidence of a non-cardiac condition as the cause of cardiac arrest. We defined PCA operationally as a sudden pulseless condition in the absence of a known non-cardiac condition to account for cardiac arrest.

The controls were selected by random digit dialling from the same population and were matched to cases on gender and age (within 7 years). For cases and controls to be eligible they were required to be between 25 and 74 years of age, free of clinically diagnosed heart disease or other life-threatening conditions, such as cancer, liver disease, lung disease or end stage renal disease, and to be married. The last condition was included as one of the eligibility criteria because PCA has a case fatality rate of greater than 80% and therefore we relied on information from surrogate respondents, spouses, to ascertain exposures to risk factors. Since the study focused on dietary intake of n-3 polyunsaturated fatty acids, we also excluded case and control subjects who might have been taking fish-oil supplements.

To ascertain exposure among cases, surviving cases and their spouses and the spouses of non-surviving cases were interviewed. For controls, both the control subjects and their spouses were interviewed. A detailed quantitative food frequency questionnaire, the seafood intake scale, was administered to determine the number of portions and the size of portions for each type of seafood. The questionnaire included a list of 35 types of seafood (25 fishes and 10 shellfishes) that are available in the Pacific Northwest. During the interview, food models were used to assess the size of portions. The spouses were asked to provide estimates of the subject's intake. Using the data published by the United States Department of Agriculture on the content (grams) of EPA and DHA per 100 grams, we computed the dietary intake of these fatty acids (in grams) over the prior month. Blood was drawn from the cases in the field at the time of the event, after resuscitation by the paramedics or soon after they had died in the field. Usually, blood was drawn within 30–45 min of the cardiac arrest. Blood from controls was drawn at the time of their interview. The blood specimens were analysed to determine the content of EPA and DHA in the red cell membranes as a percentage of the total fatty acids (Siscovick *et al*., 1995). The data set contained information on 266 cases and 356 controls. Of the 266 cases, 89 had two matched controls per case and one case had three matched controls. Table 1 provides the means, the standard deviations and the number of the observed values for the three exposure measures and other risk factors among case and control subjects. Most controls have all three measures and many cases have spouse and blood measures but only a few case subjects could provide information as many of them died in the field. Also the paramedics could not draw blood in some instances. Fig. 1 provides three scatterplots on a log-scale describing the relationships between the three measures:

(a) subjects' dietary intake,
(b) spouses' estimates of subjects' intake and
(c) the red cell membrane value.

From these scatterplots, linear regression models on a log-scale should be adequate amd hence all those measured are assumed to be in the log-scale in the subsequent development.

3. Model assumptions

The model assumptions are developed in two stages. First, we describe the model assumptions that relate the measured exposure variables to the actual dietary intake and then we describe the model assumptions relating the actual dietary intake and the disease outcome.

Table 1. Means and standard deviations SD of risk factors for PCA among the cases and controls

Characteristic	*Cases*			*Controls*		
	n	*Mean*	*SD*	*n*	*Mean*	*SD*
Dietary intake						
Subject	42	4.70	5.28	332	4.94	4.58
Spouse	266	3.76	4.97	356	4.62	4.66
Red cell value	80	4.34	1.37	292	4.74	1.13
Age	266	60	10	356	58	11
Female sex (%)	266	18.7		356	20.8	
Hypertensives (%)	266	28.4		356	15.7	
Diabetes (%)	266	12.8		356	2.8	
Former smokers (%)	266	36.3		356	43.0	
Current smokers (%)	266	36.7		356	9.3	
Fat index score	266	22	4	356	21	4
Family history of myocardial infarction or PCA (%)	266	48.4		356	46.5	
At least high school education (%)	266	61.9		356	79.7	

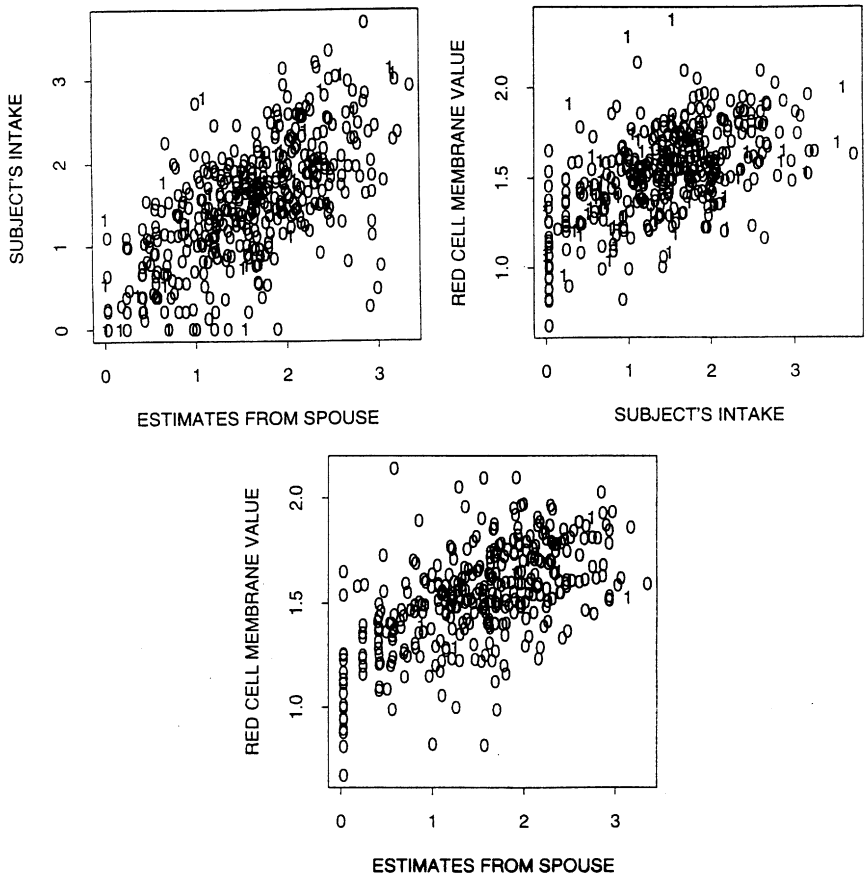

Fig. 1. Scatterplots of the three measures of dietary intake of n-3 fatty acids on the log-scale: 1, values for case subjects; 0, values for control subjects

There are several considerations in specifying the model assumptions relating the actual and reported dietary intake. First, the measurement error properties due to recall biases and response errors can be different for cases and controls. Similarly, owing to differences in their metabolic effects, the cases and controls may have different relationships between the actual dietary intake and the red cell values. Thus, we prefer to model these relationships separately for cases and controls. Second, there may be some covariates that predict or modify the reliability of the measured intake that may not be of interest in the actual case–control analysis as confounding variables. For example, the number of years that the subject and spouse were married may affect the reliability but it may not confound the relationship between the actual dietary intake and disease. Finally, the model specifications that result in the closed form expression for the conditional distributions in Gibbs sampling are desirable from the computational point of view.

3.1. Measurement error model

Suppose that T_{id} denotes the unobserved true or actual value of dietary intake of n-3 polyunsaturated fatty acids for subject $i = 1, 2, \ldots, n_d$ with disease status $d = 0, 1$. Let X_{id}, Y_{id} and Z_{id} denote the corresponding values (on a log-scale) derived from the subject's questionnaire, the spouse's questionnaire and the red cell membrane value respectively. Let w denote a vector of covariates.

The basic strategy is to model the joint distribution of (T, X, Y, Z) given w for cases and controls separately and from which the predictive distribution of T given X, Y, Z and w is constructed. The draws are obtained from the predictive distribution and are then used in the second stage of the analysis. Ignoring the subscripts for brevity and using the notation $[A|B]$ for the conditional distribution of A given B, we use the decomposition

$$\begin{aligned}[T, X, Y, Z|\Omega, w, D] &= [X, Y|T, Z, \Omega, w, D][T|Z, \Omega, w, D][Z|\Omega, w, D] \\ &= [X|T, \Omega, w, D][Y|T, \Omega, w, D][T|Z, \Omega, w, D][Z|\Omega, w, D]\end{aligned}$$

to model the joint distribution (T, X, Y, Z) given w and D, the disease status, where Ω is a vector of unknown parameters. Though there are several other possible decompositions that could be used to model this joint distribution, we chose this approach because of the apparent transparency of the model assumptions as well as the computational ease. The last equality also implies that we are assuming that, conditional on the actual exposure value T and the covariate w, the red cell membrane value provides no additional information about X and Y and the measurement errors in X and Y are independently distributed.

To describe the relationship between the reported dietary intake and the actual intake, $[X, Y|T, D, \Omega]$, we posit the regression models

$$\begin{aligned}X_{id} &= \alpha_{0d} + \alpha_{1d} T_{id} + \alpha_{2d}^{\mathrm{T}} w_{id} + e_{id}, \\ Y_{id} &= \beta_{0d} + \beta_{1d} T_{id} + \beta_{2d}^{\mathrm{T}} w_{id} + f_{id},\end{aligned} \tag{1}$$

where w_{id} is a p-dimensional vector of covariates, $(\alpha_{0d}, \alpha_{1d}, \alpha_{2d}^{\mathrm{T}}, \beta_{0d}, \beta_{1d}, \beta_{2d}^{\mathrm{T}}; d = 0, 1)$ are the regression coefficients and (e_{id}, f_{id}) are the error terms that are assumed to be mutually independent (and also independent of T_{id}) normal random variables with mean 0 and variances σ_{ed}^2 and $\kappa_d^2 \sigma_{ed}^2$ respectively.

Equations (1) describe the measurement error model where the regression coefficients define the extent of response bias or a systematic overestimation or underestimation by the subject or spouse and the effect of covariates w on this relationship. Note that these relationships are assumed to be different for cases and controls, acknowledging the possibility of recall and

response bias properties to be different in these two populations. The variance terms σ^2_{ed} and κ^2_d define the extent of random measurement error.

Next, we describe the model relating the actual dietary intake and the red cell membrane value $[T|Z, w, D, \Omega]$. Though a regression model similar to equations (1) with Z_{id} as the dependent variable can be used, we are expressing this relationship slightly differently. Any amount of n-3 fatty acids in the red cell membrane indicates that the subject has a certain dietary intake. Further, the red cell value being measured as a percentage of all fatty acids cannot be considered as an unbiased measure of the actual dietary intake. Thus we interpret the role of information provided by the red cell value as that of providing means of forming a prior distribution of the actual dietary intake. To quantify this prior information, we posit the regression model

$$T_{id} = \gamma_{0d} + \gamma_{1d} Z_{id} + \gamma^{\mathrm{T}}_{2d} w_i + g_{id} \tag{2}$$

where $(\gamma_{0d}, \gamma_{1d}, \gamma^{\mathrm{T}}_{2d})$ are the regression coefficients and the g_{id} are independent (also independent of e_{id}, f_{id} and Z_{id}) normal random variables with mean 0 and variance σ^2_{gd}. Finally, to complete the model specification, we assume that the Z_{id} are normally distributed with mean $\mu_{0d} + \mu^{\mathrm{T}}_{1d} w_i$ and variance σ^2_{zd}.

However, the models as defined above are not identifiable. There are 11 parameters for each disease group $(\alpha_{0d}, \alpha_{1d}, \beta_{0d}, \beta_{1d}, \gamma_{0d}, \gamma_{2d}, \sigma_{ed}, \kappa_d, \sigma_{gd}, \mu_{0d}, \sigma_{zd})$ other than the regression coefficients for w that essentially have to be estimated using the nine sufficient statistics based on the conditional trivariate normal distribution of (X, Y, Z) given w. Consequently, two constraints are needed to make the model identifiable. For this, we assume that the two intercepts α_{0d} and β_{0d} are equal and we denote the common value by α_d. The implication of this constraint is that when the actual intake is 0 then the subjects and spouses with comparable w will agree with each other except for some random errors. The basis of this assumption was an auxiliary study where the subjects and spouses were asked to keep food records and these were compared with the estimate based on the questionnaire. On the basis of the analysis of the auxiliary data, this assumption was deemed reasonable. We also fixed κ_d and used various values to index sensitivity analyses as described in Section 5. There are other ways in which the model can be made identifiable. For instance, a proper non-diffuse prior may be used on these parameters or we assume that one of the measures is unbiased for the actual dietary intake. However, we chose the constraints because of their empirical origins based on the auxiliary study.

The regression coefficients for w, $(\alpha_{2d}, \beta_{2d})$, are different for subjects and spouses in the model specification given in equations (1). We performed preliminary analyses to investigate the interaction between w and the disease status that is inherently implied in the model. On the basis of this preliminary investigation, we concluded that further parsimony can be achieved by assuming $\beta_{2d} = \alpha_{2d} = \delta_d$, i.e. the effect of covariates w on both subjects' and spouses' responses is the same but differs by the disease status. This assumption is similar in spirit to the analysis-of-covariance model.

We also specify a prior distribution for the unknown parameters

$$\Omega = (\alpha_d, \alpha_{1d}, \beta_{1d}, \delta_d, \sigma^2_{ed}, \gamma_{0d}, \gamma_{1d}, \mu_{0d}, \mu^{\mathrm{T}}_{1d}, \sigma^2_{zd}, \gamma^{\mathrm{T}}_{2d}, \sigma^2_{gd};\ d = 0, 1)$$

to be

$$\text{prior} \propto \prod_{d=0}^{1} \sigma^{-1}_{zd} \sigma^{-1}_{ed} \sigma^{-1}_{gd}.$$

3.2. Analysis model

If we had observed the actual dietary intake then the ultimate analysis of the case–control study would be based on fitting the logistic regression model

$$\text{logit}\{\Pr(D_{ih} = 1)\} = \theta_{0h} + \theta_1 T_{ih} + U_{ih}^{\mathrm{T}}\theta_2 \tag{3}$$

where $D_{ih} = 1$ if the subject I in the matched subjects' h (stratum) is a case subject and $D_{ih} = 0$ if he or she is a control subject, T_{ih} is the actual dietary intake, U_{ih} is a vector of confounding variables and $(\theta_{0h},\ h = 0,\ 1,\ \ldots,\ n_1,\ \theta_1,\ \theta_2^{\mathrm{T}})^{\mathrm{T}}$ is a vector of regression coefficients. The likelihood-based approach is to eliminate the nuisance parameters, the stratum-specific intercepts θ_{0h}, through a conditioning argument (Breslow and Day, 1980), i.e. to estimate $\theta = (\theta_1,\ \theta_2^{\mathrm{T}})^{\mathrm{T}}$ by maximizing the conditional likelihood

$$L = \prod_{h=1}^{n_1} \frac{1}{1 + \sum_{j=1}^{n_{0h}} \exp(\theta_1 \Delta T_j^h + \theta_2^{\mathrm{T}} \Delta U_j^h)} \tag{4}$$

where n_1 is the number of cases, n_{0h} is the number of matched controls for case h, ΔT_j^h and ΔU_j^h are respectively the differences in the exposure variable and the confounding variables between the case subject h and his or her jth matched control subject.

4. Estimation

From a frequentist perspective, the two essential quantities for inferential purposes are $\hat{\theta}(T)$, the estimate of $\theta = (\theta_1,\ \theta_2^{\mathrm{T}})$ conditional on T, and $V(T)$, the estimate of the variance of $\hat{\theta}(T)$. Since T is not observed, we propose that inference be made using the posterior mean and the variance of $\hat{\theta}(T)$,

$$\hat{\theta}^* = E\{\hat{\theta}(T)|X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}},\ W,\ D\}$$

and

$$V^* = E\{V(T)|X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}},\ W,\ D\} + V\{\hat{\theta}(T)|X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}},\ W,\ D\} \tag{5}$$

where $(X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}})$ is the observed portion of $(X,\ Y,\ Z)$ and the expectations are with respect to the predictive distribution of T given $(X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}},\ W,\ D)$.

These quantities can also be interpreted as the posterior mean,

$$E(\theta|\text{obs}) = E\{E(\theta|\text{obs},\ T)|\text{obs}\},$$

and V^* as the posterior variance,

$$V(\theta|\text{obs}) = E\{\text{var}(\theta|\text{obs},\ T)|\text{obs}\} + \text{var}\{E(\theta|T,\ \text{obs})|\text{obs}\},$$

where $\text{obs} = \{X_{\text{obs}},\ Y_{\text{obs}},\ Z_{\text{obs}},\ U,\ W,\ D\}$ is the observed data.

A fully Bayesian version that acknowledges matching of cases and controls will typically treat the stratum-specific intercepts θ_{0h} in the logistic model (1) as independent random effects with possibly a normal distribution with a common mean and variance. With a prior specified for this mean and variance and for θ, the marginal posterior distribution of θ_1 could be constructed or at least approximated. The essential difference between the hybrid approach and the fully Bayesian approach lies in the nature of the approximation of $E(\theta|\text{obs},\ T)$ and $\text{var}(\theta|\text{obs},\ T)$, i.e. the conditional likelihood function of θ given in equation (4) can be interpreted as its marginal posterior density given T, U and D (apart from the constant of proportionality independent of θ) under a uniform prior for θ. This marginal posterior density of θ given T, U and D is approximated by a multivariate normal density with mean $\hat{\theta}(T)$ and covariance matrix $V(T)$. Thus for large samples $\hat{\theta}(T)$ is an approximation for $E(\theta|T,\ \text{obs})$ and $V(T)$ for $\text{var}(\theta|T,\ \text{obs})$ under the model assumptions stated.

An alternative would be to sample directly from the density of θ given in equation (4), which is computationally difficult when the number of matching controls per case varies considerably and for a large number of covariates. An advantage of the hybrid approach, from a robustness point of view, is that the large sample approximation does not require the normality assumptions for the stratum-specific intercepts θ_{0h}, though prior investigations have shown that in the fully Bayesian approach the results are not sensitive to departures from the assumed normality (Raghunathan and Ii, 1993).

4.1. Maximum likelihood estimation

If there were no missing data in X, Y and Z, we could obtain the maximum likelihood estimates of the parameters Ω (with restricted maximum likelihood estimates of the variance components) by using the EM algorithm (Dempster *et al.*, 1977) and hence empirical Bayes estimates $\hat{T}^*$ of the actual exposure measures could be obtained. The estimate $\hat{\theta}^*$ could be approximated by $\hat{\theta}(\hat{T}^*)$.

Specifically, in the EM set-up, the complete data are (X, Y, Z, T) and the observed data are (X, Y, Z). Given the complete data, it is straightforward to show that the maximum likelihood estimates (the M-step), for $d = 0, 1$, are as follows.

(a) $\hat{\mu}_d = (W_d^{\mathrm{T}} W_d)^{-1} W_d^{\mathrm{T}} Z_d$ and $\hat{\sigma}^2_{zd} = (Z_d - W_d\hat{\mu}_d)^{\mathrm{T}}(Z_d - W_d\hat{\mu}_d)/(n_d - p - 1)$ where $\mu_d = (\mu_{0d}, \mu_{1d}^{\mathrm{T}})^{\mathrm{T}}$, $Z_d = (Z_{1d}, Z_{2d}, \ldots, Z_{n_d d})^{\mathrm{T}}$, $W_d = (\mathbf{1}, w_d)$, $w_d = (w_{1d}, w_{2d}, \ldots, w_{n_d d})^{\mathrm{T}}$, $\mathbf{1}$ is a vector of 1s and n_d is the number of individuals with disease status d.

(b) Letting $\gamma_d = (\gamma_{0d}, \gamma_{1d}, \gamma_{2d})^{\mathrm{T}}$,

$$\hat{\gamma}_d = (V_d^{\mathrm{T}} V_d)^{-1} V_d^{\mathrm{T}} T_d$$

and

$$\hat{\sigma}^2_{gd} = \frac{(T_d - V_d\hat{\gamma}_d)^{\mathrm{T}}(T_d - V_d\hat{\gamma}_d)}{n_d - p - 2}$$

where $V_d = (\mathbf{1}, Z_d, W_d)$ and $T_d = (T_{1d}, T_{2d}, \ldots, T_{n_d d})^{\mathrm{T}}$.

(c) Let $X_d = (X_{1d}, X_{2d}, \ldots, X_{n_d d})$, $Y_d = (Y_{1d}, Y_{2d}, \ldots, Y_{n_d d})$, $\phi_d = (\alpha_d, \alpha_{1d}, \beta_{1d}, \delta_d^{\mathrm{T}})^{\mathrm{T}}$, $Q_d = (X_d^2, Y_d^{\mathrm{T}})^{\mathrm{T}}$ and

$$U_d = \begin{pmatrix} \mathbf{1} & T_d & 0 & W_d \\ \mathbf{1} & 0 & T_d & W_d \end{pmatrix}.$$

It is easy to show that

$$\hat{\phi}_d = (U_d^{\mathrm{T}} M_d^{-1} U_d)^{-1} U_d^{\mathrm{T}} M_d^{-1} Q_d$$

where

$$M_d = \mathrm{diag}(1, 1, \ldots, 1, \kappa_d^2, \kappa_d^2, \ldots, \kappa_d^2)$$

is a $2n_d \times 2n_d$ diagonal weight matrix due to the different precisions of the two dietary measures and

$$\hat{\sigma}^2_{ed} = A_d/(2n_d - p - 3)$$

where

$$A_d = \sum_i \left\{ (X_{id} - \hat{\alpha}_d - \hat{\alpha}_{1d} T_{id} - \hat{\delta}_d^{\mathrm{T}} w_{id})^2 + \frac{1}{\kappa_d^2}(Y_{id} - \hat{\alpha}_d - \hat{\beta}_{1d} T_{id} - \hat{\delta}_d^{\mathrm{T}} w_{id})^2 \right\}.$$

The estimates of μ_d and σ_{zd} in (a) do not involve any unknown quantities. The unknown sufficient statistics in the remaining expressions involve T_{id} and T_{id}^2. Thus, at the E-step we need to compute the expected values of T_{id} and T_{id}^2 conditionally on the observed values and the current estimate of the parameters. Given the parameter estimates and the observed data, it can be shown that the T_{id} are independent normals with means

$$\hat{T}_{id}^* = \left\{ \frac{\hat{\alpha}_{1d}^2 + \hat{\beta}_{1d}^2/\kappa_d^2}{\hat{\sigma}_{ed}^2} + \frac{1}{\hat{\sigma}_{gd}^2} \right\}^{-1} \left\{ \frac{(X_{id} - \hat{\alpha}_d - \hat{\delta}_d^{\mathrm{T}} w_{id})\hat{\alpha}_{1d} + (Y_{id} - \hat{\alpha}_d - \hat{\delta}_d^{\mathrm{T}} w_{id})\hat{\beta}_{1d}/\kappa_d^2}{\hat{\sigma}_{ed}^2} + \frac{\hat{\gamma}_{0d} + \hat{\gamma}_{1d} Z_{id} + \hat{\gamma}_{2d}^{\mathrm{T}} w_{id}}{\hat{\sigma}_{gd}^2} \right\} \tag{6a}$$

and common variance

$$\left(\frac{\hat{\alpha}_{1d}^2 + \hat{\beta}_{1d}^2/\kappa_d^2}{\hat{\sigma}_{ed}^2} + \frac{1}{\hat{\sigma}_{gd}^2} \right)^{-1}. \tag{6b}$$

Thus, starting with an initial guess for the unknown parameters, the EM algorithm iterates between the E-step based on equations (6) and the M-steps until the parameter estimates stabilize.

At the convergence of the EM iterations, $\hat{T}_{id}^*$ are the empirical Bayes estimates of T_{id} which can be substituted for T in the logistic model. The estimates θ may be obtained again by maximizing the conditional likelihood with $\hat{T}$ and U as the covariates. However, the asymptotic variance $V(\hat{T})$ will be an underestimate of the true sampling distribution as it ignores the uncertainty in not knowing T. Either the bootstrap (Laird and Louis, 1987) or the jackknife (Raghunathan, 1993) approach specifically discussed in the context of empirical Bayes analysis will have to be used to reflect the increased uncertainty in the point estimate of θ.

4.2. Gibbs sampling

A straightforward method for computing $\hat{\theta}^*$ and V^* defined in equations (4) when some values of X, Y and Z are missing is through simulation techniques such as Gibbs sampling (Gelfand and Smith, 1990) which has received considerable attention in the recent literature. Briefly, Gibbs sampling in the present example involves drawing values from $\Pr(T|X_{\mathrm{obs}}, Y_{\mathrm{obs}}, Z_{\mathrm{obs}}, w, D)$ or equivalently drawing values from $\Pr(T, X_{\mathrm{mis}}, Y_{\mathrm{mis}}, Z_{\mathrm{mis}}, \Omega|X_{\mathrm{obs}}, Y_{\mathrm{obs}}, Z_{\mathrm{obs}}, w, D)$ where

$$\Omega = (\alpha_d, \alpha_{1d}, \beta_{1d}, \delta_d, \sigma_{ed}, \gamma_{0d}, \gamma_{1d}, \gamma_{2d}, \sigma_{gd}, \mu_{zd}, \sigma_{zd};\ d = 0, 1)$$

by drawing from each univariate conditional distribution (or that of a subvector) of the missing value or the parameter in a cyclic fashion each time replacing the old values by the most recently drawn values. It is preferable to ignore the initial few cycles to eliminate the effect of the starting values. Also, Gelman and Rubin (1992) suggested using replicates or parallel cycles with different starting values to eliminate their effects further.

Let t_0 be the number of initial cycles that are ignored to eliminate the effect of the starting values. For $t = t_0, t_0 + 1, \ldots, t_0 + N$, let $T^{(t)}$ denote the value of T drawn in the tth Gibbs cycle. Let $\hat{\theta}^{(t)} = \hat{\theta}(T^{(t)})$ denote the estimate θ from the logistic model using $T^{(t)}$ instead of T and let $v^{(t)} = V(T^{(t)})$ denote the corresponding asymptotic variance of $\hat{\theta}^{(t)}$. Using the ergodic results (Gelfand and Smith, 1990) it can be shown that

$$\theta^* \approx \sum_{t=t_0}^{t_0+N} \hat{\theta}^{(t)}/N$$

and

$$V^* \approx \sum_{t=t_0}^{t_0+N} v^{(t)}/N + \sum_{t=t_0}^{t_0+N} (\hat{\theta}^{(t)} - \theta^*)(\hat{\theta}^{(t)} - \theta^*)^{\mathrm{T}}/(N-1).$$

We now briefly describe the essential Gibbs sampling steps. For notational convenience we shall use 'Rest' to designate the rest of the variables and parameters other than the argument of the density. From the model specification for $(T, X, Y, Z|w, D, \Omega)$ and the prior distribution described in Section 3, the following results are easily derived.

(a) $\sigma_{zd}^{-2}\Sigma_i (Z_d - W_d^{\mathrm{T}}\hat{\mu}_d)^{\mathrm{T}}(Z_d - W_d^{\mathrm{T}}\hat{\mu}_d)|\text{Rest} \sim \chi^2_{n_d - p - 1}$ where Z_d and W_d are defined in Section 4.1 and $\hat{\mu}_d = (W_d^{\mathrm{T}}W_d)^{-1}W_d^{\mathrm{T}}Z_d$.
(b) $\mu_{zd}|\text{Rest} \sim \text{normal}\{\hat{\mu}_d, \sigma_{zd}^2(W_d^{\mathrm{T}}W_d)^{-1}\}$.
(c) $\sigma_{gd}^{-2}(T_d - V_d\hat{\gamma}_d)^{\mathrm{T}}(T_d - V_d\hat{\gamma}_d)|\text{Rest} \sim \chi^2_{n_d - p - 2}$ where $\gamma_d = (\gamma_{0d}, \gamma_{1d}, \gamma_{2d})^{\mathrm{T}}$, $\hat{\gamma}_d = (V_d^{\mathrm{T}}V_d)^{-1} \times V_d^{\mathrm{T}}T_d$ and T_d and V_d are defined in Section 4.1.
(d) $\gamma_d|\text{Rest} \sim \text{normal}\{\hat{\gamma}_d, \sigma_{gd}^2(V_d^{\mathrm{T}}V_d)^{-1}\}$.
(e) Let $\hat{\phi}_d = (U_d^{\mathrm{T}}M_d^{-1}U_d)^{-1}U_d^{\mathrm{T}}M_d^{-1}Q_d$ where ϕ_d, Q_d, U_d and M_d are defined in Section 4.1. $\sigma_{ed}^{-2}A_d|\text{Rest} \sim \chi^2_{2n_d - p - 3}$ where

$$A_d = \sum_i \left\{ (X_{id} - \hat{\alpha}_d - \hat{\alpha}_{1d}T_{id} - \hat{\delta}_d^{\mathrm{T}}w_{id})^2 + \frac{1}{\kappa_d^2}(Y_{id} - \hat{\alpha}_d - \hat{\beta}_{1d}T_{id} - \hat{\delta}_d^{\mathrm{T}}w_{id})^2 \right\}.$$

(f) $\phi_d|\text{Rest} \sim \text{normal}\{\hat{\phi}_d, \sigma_{ed}^2(U_d^{\mathrm{T}}M_d^{-1}U_d)^{-1}\}$.
(g) $T_{id}|\text{Rest}$ for $i = 1, 2, \ldots, n_d$, $d = 0, 1$, are independent normal distributions with mean and variance given in equations (6) except that they are evaluated at the current drawn value of the parameters.
(h) $X_{id}|\text{Rest}$ for $i = 1, 2, \ldots, n_d$, $d = 0, 1$, are independent normal distributions with means $\alpha_d + \alpha_{1d}T_{id} + \delta_d^{\mathrm{T}}w_{id}$ and common variance σ_{ed}^2.
(i) $Y_{id}|\text{Rest}$ for $i = 1, 2, \ldots, n_d$, $d = 0, 1$, are independent normal distributions with means $\alpha_d + \beta_{1d}T_{id} + \delta_d^{\mathrm{T}}w_{id}$ and common variance $\kappa_d^2\sigma_{ed}^2$.
(j) $Z_{id}|\text{Rest}$ for $i = 1, 2, \ldots, n_d$, $d = 0, 1$, are independent normal distributions with means

$$(\gamma_{1d}^2/\sigma_{gd}^2 + 1/\sigma_{zd}^2)^{-1}\{(T_{id} - \gamma_{0d} - \gamma_{2d}^{\mathrm{T}}w_{id})\gamma_{1d}/\sigma_{gd}^2 + (\mu_{0d} + \mu_{1d}w_{id})/\sigma_{zd}^2\}$$

and common variance $(\gamma_{1d}^2/\sigma_{gd}^2 + 1/\sigma_{zd}^2)^{-1}$.

Thus Gibbs sampling involves first drawing the initial values for the missing components of (X, Y, Z), T and the parameters Ω and then using the conditional distributions given above to update the drawn values sequentially.

The initial values can be obtained as follows. First, the missing values in (X, Y, Z) can be drawn from the appropriate conditional distributions derived from a trivariate normal regression model with w as the core dependent variable with the parameters estimated on the basis of the complete cases (see, for example, Box and Tiao (1973)). For example, suppose that, say, X is missing and Y and Z are observed; then the missing X can be drawn from a conditional normal distribution of X given Y, Z and w. Once the missing values in (X, Y, Z) have been filled in, the parameter estimates may be obtained through the maximum likelihood approach discussed in Section 4, and then conditionally on these parameter estimates the values of T may be drawn on the basis of the conditional distribution given in item (g) above.

5. Analysis of example

For the example described in Section 2, we included all the covariates listed in Table 1 in *w*. We also included number of years married, spouse's education and the occupation of the subjects as measured by a dummy variable 'working away from home' (yes $\equiv$ 1; no $\equiv$ 0) in the measurement error model given in equation (1). All the continuous covariates were centred at the mean value for the combined sample of cases and controls, so that the intercepts are interpretable as the estimated response bias for an 'average' member of our sample.

Using the method described in the previous section we obtained the estimate θ^* of θ and the associated covariance matrix V^*. We used the conditional logistic regression approach because of the matched design. In the Gibbs sampling we ignored the first 10 000 (t_0) draws to eliminate the effect of starting values and 50 000 (N) draws were used to construct the point estimates and their standard errors. The Gibbs draws were obtained in 10 replicates of 5000 draws in each.

On the basis of a preliminary analysis the linear and quadratic function of the drawn values of the actual dietary intake and other confounding variables listed in Table 1 were used as independent variables in the logistic model. The point estimates and their standard errors are given in the second and the third columns of Table 2 for $\kappa_d^2 = 1$, $d = 0, 1$. The next pair of columns provide the point estimates and their standard errors when the spouses' data are used as a substitute for the actual dietary intake. Although the estimates of the regression coefficients for the confounding variables do not change much, the adjusted log-odds ratio is attenuated towards 0 when the spouses' data are used instead of the estimated actual dietary intake. Thus it seems that the protective effect of dietary intake on the incidence of PCA is underestimated using the proxy data. To explore the sensitivity, we used two other values, $\kappa_d^2 = 1.5$ and $\kappa_d^2 = 2$. The estimates and their standard errors under these alternative values of κ_d^2 are also provided in Table 2. The point estimates do not change much but the standard errors are larger as expected because of the imprecision that is inherent in a spouse's estimate of a subject's intake.

6. Simulation study

The method described in this paper uses a combination of Bayesian and frequentist ideas. For routine applications, however, it is desirable to investigate the frequency properties of the estimates obtained by using this method. We therefore conducted a simulation study to investigate the

Table 2. Estimated regression coefficients (and their standard errors SE) in the logistic model

Variable	*Bayes,* $\kappa_d^2 = 1$		*Proxy (spouse) data*		*Bayes,* $\kappa_d^2 = 1.5$		*Bayes,* $\kappa_d^2 = 2$	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Dietary intake								
Linear	−0.1961	0.0647	−0.1193	0.0433	−0.1817	0.0797	−0.1820	0.0824
Quadratic	0.0079	0.0023	0.0038	0.0015	0.0068	0.0024	0.0068	0.0029
Age	0.1193	0.0174	0.1210	0.0370	0.1199	0.0174	0.1199	0.0172
Current smoker	1.6485	0.3770	1.8828	0.3233	1.6489	0.3771	1.6489	0.3771
Former smoker	0.3338	0.2419	0.3237	0.2496	0.3336	0.2422	0.3336	0.2443
Family history of myocardial infarction or sudden death	0.7085	0.2005	0.7044	0.2189	0.7086	0.2011	0.7088	0.2017
Fat index score	0.0211	0.0282	0.0242	0.0276	0.0217	0.0284	0.0217	0.0292
Hypertension	0.4489	0.2732	0.4501	0.2623	0.4489	0.2741	0.4489	0.2472
Diabetes	1.6066	0.4245	1.6055	0.4135	1.6069	0.4244	1.6068	0.4244
More than high school education	−0.5992	0.2458	−0.6006	0.2449	−0.5997	0.2457	−0.5996	0.0248

bias and the mean-square error of the point estimate and the exact coverage of the nominal 95% confidence interval of the adjusted log-odds ratio θ_1 derived by using the method described in this paper.

We considered two simulation conditions. In the first, we generated the data T, X, Y and Z on 1000 individuals under the assumptions stated in Section 2. We used the estimates of α_d, α_{1d}, β_{1d}, σ_{ed}, σ_{gd}, μ_{zd} and σ_{zd}, $d = 0$, 1, obtained in the example which are given in Table 3 and κ_d^2 was fixed at 1.

We first generated D such that roughly 43% of the Ds were 1 as in the example discussed in the previous section. Next we generated values from the distributions $[Z|D]$, $[T|Z, D]$, $[X|T, D]$ and $[Y|T, D]$. This approximately resulted in the true log-odds ratio of -0.97 for a one-unit increase in T.

From the complete data we created incomplete data on X and Z that were very similar to the numbers given in Table 1. Thus, this simulation condition is replicating data that are similar to the example data. The second simulation condition was similar to the first except that we fixed $\alpha_d = 0$ and $\alpha_{1d} = \beta_{1d} = 1$, i.e. both X and Y are unbiased for the actual value T. In both simulation conditions there were no other covariates w.

For each simulation condition, 10 000 data sets were generated and the point and interval estimates of θ_1 were obtained. The first 2000 cycles were ignored in the Gibbs sequence for each data set and the posterior mean and variance were computed on the basis of the next 5000 draws. Several other choices of t_0 and N on a smaller set of simulated data sets resulted in similar results. All computations were performed on a SUN SPARCstation 20 using GAUSS programming language (Aptech Systems, 1992).

Table 4 provides the bias and the mean-square error of the point estimate and the exact confidence coverage of the nominal 95% confidence interval. For comparison, Table 4 also provides the same quantities except using Y (as in the spouse's estimate which has no missing values) as a substitute for the actual value T. The estimates based on the procedure described in this paper are almost unbiased and the confidence intervals are well calibrated. In contrast, pretending that Y is the actual value leads to severely biased estimates and poorly calibrated confidence intervals even for the second simulation condition.

We wanted to explore the sensitivity of the sampling properties to the normality of the distributions of X, Y, Z and T. Therefore, we repeated the simulation study just described except that the values of X, Y, Z and T were drawn from a scaled χ^2-distribution with 4 degrees of freedom. The scaling ensured that the means, variances and covariances were the same as in the

Table 3. Parameter estimates of the measurement error model for the example when $\kappa_0^2 = \kappa_1^2 = 1$

Parameter	*Case estimate* ($d = 1$)	*Control estimate* ($d = 0$)
α_d	−3.2245	−2.8871
α_{1d}	1.1178	1.8963
β_{1d}	1.1923	1.8699
σ_{ed}^2	0.0732	0.1699
γ_{0d}	0.9931	0.8454
γ_{1d}	0.7654	0.9221
σ_{gd}^2	0.0431	0.0721
μ_{0d}	1.5667	1.5263
σ_{zd}^2	0.1827	0.2449

Table 4. Bias and the mean-square error of the point estimates and the exact coverage of the interval estimates of the adjusted log-odds ratio obtained by using the approximate Bayes and naïve methods under various simulation conditions

Distribution	*Simulation*	*Bias*		*Mean-squared error*		*Coverage*	
		Bayes	*Naïve*	*Naïve*	*Bayes*	*Bayes*	*Naïve*
Normal	1	0.0034	0.6221	0.1265	0.4248	96	1
Normal	2	0.0021	0.3212	0.1007	0.3990	94	22
χ^2	1	0.0054	0.8337	0.1562	0.5287	92	3
χ^2	2	0.0062	0.8339	0.2106	0.8978	94	21

normal case. The results are given in the bottom half of Table 4. There is a modest increase in both the bias and the mean-square error and the confidence interval becomes somewhat anti-conservative.

7. Discussion

In this paper we have proposed, and evaluated, an analytical strategy for analysing data from a case–control study where the exposure variable is measured by using multiple measurement sources none of which are unbiased for the actual exposure variable. We have developed a hybrid approach where a Bayesian model is used to estimate the actual dietary intake based on the mismeasured covariates and a likelihood-based approach is used to estimate the relative risk of PCA with respect to the actual dietary intake. Our approach incorporates the increase in uncertainty due to using an estimated rather than the actual dietary intake while fitting the ultimate logistic regression model. However, for large samples the estimates of the log-odds ratios may also be viewed as an approximate posterior mean under a fully Bayesian model. We have also developed a maximum likelihood approach to fit the measurement error model when there are no missing values in the mismeasured covariates.

The particular data set that was analysed suggests that relying on the spouses' data as a marker for the actual dietary intake can underestimate the protective effect of dietary intake of n-3 polyunsaturated fatty acids. For example, the adjusted odds ratio comparing 3.3 g of dietary intake (one fatty fish meal per week) with those who do not eat any seafood is 0.7 (the 95% confidence interval is [0.5, 0.9]) using the proxy data and the corresponding figure using the estimated actual dietary intake is 0.4 (95% confidence interval [0.3, 0.6]).

The limited simulation study suggests that the point and interval estimates also have desirable frequency properties and are fairly robust to modest departures from the assumed normality. A further detailed investigation is necessary to explore fully the robustness and sensitivity of the inference to the model assumptions.

A limitation of the approach that was taken is that it is predicated on the regression relationships between T and X, Y and Z. These are technically unverifiable assumptions in the absence of validation data that provide information at least on all the three pairs (X, T), (Y, T) and (Z, T) of relationships. Hence, it is important to apply the method by assuming different relationships and to explore the sensitivity to the regression relationships stated. For example, we performed an analysis identical with that described in Section 4 except using the original scale in all three regression models given in equations (1) and (2). The results were quite similar to those given in Table 2 and the maximum difference in the point estimates of the regression coefficient was less than 5% of the standard error.

References

Aptech Systems (1992) *The GAUSS System Version 3.1*. Maple Valley: Aptech Systems.

Armstrong, B. G., Whittemore, A. S. and Howe, G. R. (1989) Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statist. Med.*, **8**, 1151–1163.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case–Control Studies*. Lyon: International Agency for Research on Cancer.

Buonaccorsi, J. P. (1990) Double sampling for exact values in the normal discriminant model with application to binary regression. *Communs Statist. Theory Meth.*, **19**, 4569–4586.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carroll, R. J. and Stefanski, L. A. (1990) Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Am. Statist. Ass.*, **85**, 652–663.

Carroll, R. J. and Wand, M. P. (1991) Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc.* B, **53**, 573–585.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Efron, B. (1975) The efficiency of logistic regression compared to Normal discriminant analysis. *J. Am. Statist. Ass.*, **72**, 557–565.

Fuller, W. A. (1987) *Measurement Error Models*. New York: Wiley.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

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

Ii, Y. (1992) Semiparametric measurement error models. *PhD Thesis*. Department of Biostatistics, University of Washington, Seattle.

Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *J. Am. Statist. Ass.*, **82**, 739–750.

Pepe, M. S. and Fleming, T. R. (1991) A nonparametric method for dealing with mismeasured covariate data. *J. Am. Statist. Ass.*, **86**, 108–113.

Raghunathan, T. E. (1993) A quasi-empirical Bayes method for small area estimation. *J. Am. Statist. Ass.*, **88**, 1444–1448.

Raghunathan, T. E. and Ii, Y. (1993) Analysis of binary data from a multicentre clinical trial. *Biometrika*, **80**, 127–139.

Raghunathan, T. E. and Siscovick, D. S. (1996) A multiple-imputation analysis of a case–control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Appl. Statist.*, **45**, 335–352.

Richardson, S. and Gilks, W. R. (1993) A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am. J. Epidem.*, **137**, 430–442.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Ass.*, **91**, 722–732.

Rosner, B., Willet, W. C. and Spiegelman, D. (1992) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Med.*, **8**, 1051–1070.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Siscovick, D. S., Raghunathan, T. E., King, I., Weinmann, S., Wicklund, K. G., Albright, J., Bovbjerg, V., Arbogast, P., Kushi, L., Cobb, L., Copass, M. K., Psaty, B. M., Retzlaff, B., Childs, M. and Knopp, R. H. (1995) Dietary intake and cell-membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *J. Am. Med. Ass.*, **274**, 1363–1367.

Stefanski, L. A. and Carroll, R. J. (1989) Efficient scores and optimal scores in generalized linear measurement error models. *Biometrika*, **74**, 703–716.

Whittemore, A. S. (1989) Errors in variables regression using Stein estimates. *Am. Statistn*, **43**, 226–228.