Bayesian Adjustment for Exposure Misclassification in Case-Control Studies

Rong Chu

Clinical Epidemiology and Biostatistics, McMaster University

Hamilton, Ontario, L8N 3Z5, Canada

Paul Gustafson

Department of Statistics, University of British Columbia

Vancouver, British Columbia, V6T 1Z2, Canada

Nhu Le

BC Cancer Agency

Vancouver, British Columbia, V5Z 1L3, Canada

Abstract

Summary: Poor measurement of explanatory variables occurs frequently in observational studies. Error-prone observations may lead to biased estimation and loss of power in detecting the impact of explanatory variables on the response. We consider misclassified binary exposure in the context of case-control studies, assuming the availability of validation data to inform the magnitude of the misclassification. A Bayesian adjustment to correct for the misclassification is investigated. Simulation studies show that the Bayesian method can have advantages over non-Bayesian counterparts, particularly in the face of a rare exposure, small validation sample-sizes, and uncertainty about whether exposure misclassification is differential or non-differential. The method is illustrated via application to several real studies.

Keywords: Bayesian methods; case-control study; exposure misclassification; simulation-extrapolation.

1 Introduction

In biomedical studies, the misclassification problem arises when a categorical exposure variable T is not precisely recorded. Instead of T, an approximate measurement or a *surrogate*, X is obtained. Replacing T with X in data analysis without accounting for the misclassification does not generally lead to valid inference about the association between T and a health-related response Y. Hence, the goal of adjustment for mismeasurement is to achieve valid inference about the (T, Y) relationship from (X, Y) data. In this paper, we restrict ourself to misclassification problems on a binary exposure variable (T =0, 1) in case-control studies (Y = 0, 1 for controls, cases) and no other covariates at play. We consider the setting whereby a "validation subsample" is available, i.e, for the majority of subjects only (X, Y) data are obtained, but for a (randomly-selected) minority (T, X, Y) are obtained. Such a design can arise when X is inexpensive and/or quick to measure whereas T is expensive and/or time-consuming to measure. Table 1 described the data structure. While each cell a_{ij} in the validation data is fully specified ($i = 0, 1$,$j = 1, 2, 3, 4$), only margins $a_{05}, a_{06}, a_{15}, a_{16}$ in the main data are recorded.

It is sometimes sensible to assume the conditional distribution of X given T and Y does not depend on Y, which is known as *nondifferential* misclassification. In other circumstances, the sampling scheme of case-control studies (explanatory variables are retrieved after the diagnosis) may well lead to the so called *differential* measurement error, i.e. the conditional distribution of the surrogate X given the unobservable exposure T also depends on the response Y. When information about covariates is collected through some "self-report" mechanism, subjects with target clinical outcomes may tend to erroneously "blame" a set of risk factors for their conditions, or "ignore" previous exposure to avoid any connection between behaviour and disease.

There is a large literature on correcting for exposure mismeasurement, for example Barron [1], Marshall [2], Lyles [3], Carroll et al. [4]. Most work approaches the problem from a frequentist perspective, assuming complete knowledge of whether the misclassification is nondifferential or differential. The simulation extrapolation method and latent class logistic regression model were developed to tackle the same problem [5, 6]. On the other hand, the dramatic improvement of computational capability and the development of indirect simulation techniques such as Markov chain Monte Carlo (MCMC) make it possible to explore misclassification problems from a Bayesian perspective [7, 8, 9, 10]. In fact, partial prior knowledge of misclassification probabilities is often accessible to medical researchers, which makes Bayesian analysis an appealing approach.

Therefore in this paper, we primarily introduce a series of Bayesian methods suitable for different misclassification assumptions. Their performance will be closely compared to those of the maximum likelihood estimates (MLEs) and simulation extrapolation (SIMEX) method, using simulation studies and real-life examples. Section 2 presents detailed methodology for the proposed Bayesian methods. Section 3 discusses the comparative behaviours of the three methods based on simulation studies. Sections 4 and 5 present the performances of Bayesian and other methods via case-control studies with misclassified exposure variables and validation sub-samples. Section 6 provides some concluding remarks.

2 Bayesian adjustment for misclassification

Let us denote the true exposure prevalences amongst controls and cases by $r_i = P(T = 1|Y = i), i = 0, 1$. The retrospective odds ratio describing the correlation between the response and explanatory variable is defined as

$$\Phi_r = \frac{r_1/(1-r_1)}{r_0/(1-r_0)}.$$

Sensitivity (SN) and specificity (SP) jointly measure the magnitude of exposure misclassification. In the scenarios subject to *differential* misclassification, the surrogate X and the response Y are not independent, given the unobserved true exposure T. The sensitivities and specificities among cases and controls can be formulated as, $SN_i = P(X = 1|T = 1, Y = i), SP_i = P(X = 0|T = 0, Y = i)$, $i = 0, 1$. Prevalences of the *apparent* exposure for diseased and non-diseased individuals are denoted by $r_i^* = P(X = 1|Y = i) = r_i SN_i + (1 - r_i)(1 - SP_i)$, $i = 0, 1$. The degree of misclassification can also be expressed by the positive predictive value (PPV) and negative predictive value (NPV), where

$$\begin{aligned} PPV_i &= P(T = 1|X = 1, Y = i) = \frac{SN_i r_i}{SN_i r_i + (1 - SP_i)(1 - r_i)} \quad (1) \\ NPV_i &= P(T = 0|X = 0, Y = i) = \frac{SP_i(1 - r_i)}{SP_i(1 - r_i) + (1 - SN_i)r_i} \quad (2) \end{aligned}$$

It is easy to justify that, in the main study the actual number of subjects of positive exposure status (b_{i1}) amongst those who are apparently exposed in either case or control group (a_{i5}) follows a Binomial distribution, i.e. $b_{i1} \sim Binomial(a_{i5}, PPV_i)$. Similarly, conditioning on the number of cases or controls with negative apparent exposure status (a_{i6}), the number of truly unexposed subjects (b_{i4}) follows $Binomial(a_{i6}, NPV_i)$, for $i = 0, 1$.

When the nondifferential misclassification condition is fulfilled, meaning the conditional distribution of $X|T, Y$ does not depend on Y, it follows immediately that $SN_0 = SN_1 = SN$, $SP_0 = SP_1 = SP$. However it is worth pointing out that, nondifferential misclassification does not imply equality of cases and controls regarding the predictive values (PPV_i, NPV_i).

2.1 Prior distributions

The exposure prevalances r_0, r_1, sensitivities SN_0, SN_1, and specificities SP_0, SP_1 are the parameters of interest. By converting into a logit scale, $logit(x) = \log\{x/(1-x)\}$, the prior information concerning these parameters can be modeled using bivariate normal distributions [11]. The actual exposure prevalences (r_i), sensitivities(SN_i) and specificities (SP_i) of X as a surrogate for T are assumed to be uncorrelated of one another, with,

$$\begin{pmatrix} logit(r_0) \\ logit(r_1) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

$$\begin{pmatrix} logit(SN_0) \\ logit(SN_1) \end{pmatrix} \sim N\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho_2\tau_1\tau_2 \\ \rho_2\tau_1\tau_2 & \tau_2^2 \end{pmatrix}\right),$$

$$\begin{pmatrix} logit(SP_0) \\ logit(SP_1) \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \begin{pmatrix} \delta_1^2 & \rho_3\delta_1\delta_2 \\ \rho_3\delta_1\delta_2 & \delta_2^2 \end{pmatrix}\right).$$

It follows immediately that,

$$logit(SN_0) - logit(SN_1) \sim N(\nu_1 - \nu_2, \tau_1^2 + \tau_2^2 - 2\rho_2\tau_1\tau_2) \tag{3}$$

$$logit(SP_0) - logit(SP_1) \sim N(\gamma_1 - \gamma_2, \delta_1^2 + \delta_2^2 - 2\rho_3\delta_1\delta_2) \tag{4}$$

Our prior beliefs can be reflected through the *hyperparameters*, μ_i, σ_i, ν_i, τ_i, γ_i, δ_i and ρ_j. For instance, we proceed to set the prior distributions on the misclassification parameters as follows. We set $\nu_1 = \nu_2$, $\gamma_1 = \gamma_2$, $\tau_1^2 = \tau_2^2$, $\delta_1^2 = \delta_2^2$ to reflect an absence of knowledge about the "direction" of possible differentiality in the exposure assessment, with the assigned values to these quantities then reflecting prior belief about the extent of exposure misclassification. Put another way, we are expressing *exchangeable* prior beliefs about the misclassification of controls versus the misclassification of cases.

As a result, setting $\rho_2 = \rho_3 = 1$ implies that $SN_0 = SN_1$ and $SP_0 = SP_1$, which corresponds to nondifferential misclassification. Conversely, setting $\rho_2 = \rho_3 = 0$ implies independence of SN_0 and SN_1, and independence of SP_0 and SP_1. This intuitively reflects the notion that sensitivities or specificities are free to vary by themselves, and can be interpreted as "fully differential" misclassification. We will describe situations in between ($0 < \rho_j < 1$, $j = 2, 3$) as corresponding to "nearly nondifferential" misclassification, particularly when each ρ_j is close to one. This setting is useful when investigators postulate that the nondifferential assumption might hold, and that should it be violated, the extent of violation is not likely to be severe.

Similarly, we set $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ to be "unbiased", a *priori* concerning the direction of any exposure-disease association. The particular choice of values is dictated by belief about plausible values for exposure prevalence. We can then choose ρ_1 to obtain plausible prior for the effect size.

2.2 Posterior simulation

As is common in problems with "latent structure", we can implement Bayesian inference via simulation from the distribution of parameters and unobservables given observables. In the fully-differential and nearly-nondifferential cases, this amounts to sampling from the distribution of parameter $\theta = (r_0, r_1, SN_0, SN_1, SP_0, SP_1)$ and latent variables b_{ij} given observed data a_{ij}. It is easy to verify that in the related problem where the prior on θ is comprised of independent uniform distributions (or more generally independent beta distributions) for each parameter, that Gibbs sampling is possible. That is, in the related problem each component of θ has a standard "full conditional" distribution. Gibbs sampling has the nice features that (i) no tuning constants are involved, and (ii) proposed moves are always accepted. Therefore we adapt this approach to the actual problem at hand by implementing a Metropolis-Hastings algorithm, *using the full-conditionals for the related problem to generate proposals.* Thus tuning is still not needed. Moreover, the acceptance probability for each proposal will depend only on the ratio of prior densities, i.e., the specified prior based on bivariate normal distributions versus the uniform prior in the related problem. Thus we find high acceptance rates, and in general this algorithm performs well. Note also that the same computational strategy can be adopted in the nondifferential case, via the smaller parameter vector $\theta = (r_0, r_1, SN, SP)$. The Bayesian method is implemented in R and downloadable from `http://www.stat.ubc.ca/People/Home/index.php?person=gustaf`.

3 Simulation Studies

3.1 Data Simulation

In order to demonstrate the comparative performance of Bayesian adjustment against other statistical approaches, we conduct a simulation study using both relatively low $((r_0, r_1) = (0.0400, 0.0698))$ and high $((r_0, r_1) = (0.250, 0.375))$ exposure prevalences, implying an odds ratio $\Phi = \{r_1/(1-r_1)\}/\{r_0/(1-r_0)\}$ of 1.8 for the true exposure T and the response Y in either case. At each prevalence setting, four misclassification scenarios concerning different levels of differentiality are built. Under each scenario, 400 datasets each of sample size 960 are generated ($\sum_{j=1}^{4} a_{ij} = 80$, $a_{i5} + a_{i6} = 400$).

Data in scenario 1 are simulated under nondifferential misclassification, with increasing degrees of differentiality in scenarios 2, 3, and 4. To mimic the occurrence of erroneously “blaming” or “ignoring” a risk factor, we let the misclassification arising for the cases grow across scenarios, as follows.

- *Scenario 1*: (SN_0, SN_1)=(0.80, 0.80), (SP_0, SP_1)=(0.90, 0.90)
- *Scenario 2*: (SN_0, SN_1)=(0.80, 0.75), (SP_0, SP_1)=(0.90, 0.85)
- *Scenario 3*: (SN_0, SN_1)=(0.80, 0.70), (SP_0, SP_1)=(0.90, 0.80)
- *Scenario 4*: (SN_0, SN_1)=(0.80, 0.65), (SP_0, SP_1)=(0.90, 0.75)

Three Bayesian methods, adopting nondifferential, nearly nondifferential and differential prior distributions respectively, are applied to each dataset, to adjust for possible misclassifications and assess the association between the true exposure and outcome.

3.2 Choice of Hyperparameters

According to Section 2, under the assumptions that $\mu_1 = \mu_2 = \mu$, $\sigma_1 = \sigma_2 = \sigma$, $\nu_1 = \nu_2$, $\gamma_1 = \gamma_2 = \gamma$, $\tau_1 = \tau_2 = \tau$ and $\delta_1 = \delta_2 = \delta$, we assign $\mu = -1.946$, $\sigma = 0.993$ to model the prior information that the logit true exposures are normally distributed with central 95% probability between logit(0.02) and logit(0.5). Mild correlation between r_0 and r_1 ($\rho_1 = 0.3$) is selected to allow a relatively large prior standard deviation of 1.175 for $logOR$ around mean 0. Similarly, we set $\nu = \gamma = 1.675$, $\tau = \delta = 0.648$ to represent the prior knowledge that the logit sensitivity and logit specificity are normally distributed within logit(0.6) and logit(0.95) with 95% probability. As discussed in Section 2, we set $\rho_2 = \rho_3 = 1$ to reflect nondifferential misclassification; $\rho_2 = \rho_3 = 0$ to express prior belief in differential misclassification.

The choice of ρ_2, ρ_3 for nearly nondifferential misclassification requires extra work. We note that by setting $\rho_2 = \rho_3 = 0.95$ we attain:

$$P\{|logit(SN_1) - logit(SN_0)| < 0.1\} = P\{|logit(SP_1) - logit(SP_0)| < 0.1\} = 0.3746,$$

and (by simulation)

$$P\{|SN_1 - SN_0| < 0.01\} = P\{|SP_1 - SP_0| < 0.01\} = 0.32252.$$

This seems reasonable as an encapsulation of the notion that deviations from nondifferentiality are not likely severe.

3.3 Model comparison

Bayesian statistical inferences are conducted based on samples drawn from 10000 MCMC iterations, after we discard the first 1000 simulations to diminish the effect of initial distributions.

The performance of Bayesian methods is constrasted with maximum likelihood (ML) and SIMEX methods. MLEs for model parameters under differential misclassification are calculated using closed-form expressions given by Lyles Lyles [3]. A numerical optimizer (function "optim()" in **R**) is adopted to maximize the log likelihood under nondifferential misclassification. The asymptotic variance of the log odds-ratio estimator is attainable by the multivariate Delta method in these cases.

The simulation extrapolation (SIMEX) method originates in continuous measurement error settings [12]. The method introduces artificial extra measurement error to the data in question, in order to infer a relationship between the magnitude of measurement error and the estimate of the exposure-disease relationship. This relationship is then extrapolated back to the point of zero measurement error, to give an estimate which is adjusted for this error. Recently, Küchenhoff et al. extended the SIMEX procedure to the case of misclassified categorial data [5]. In brief, extra misclassification is introduced by raising the misclassification matrix to a power $\lambda > 1$, for multiple values of λ. The relationship between the point estimate of interest and λ is then extrapolated back to the $\lambda = 0$ setting of no misclassification (i.e., misclassification matrix equal to the identity matrix). The corresponding software package [13] allows different choices of extrapolation function and different methods for the calculation of standard errors. We therefore reports multiple sets of results for SIMEX. Note also that the SIMEX procedure is operationalized by "plugging in" estimates of sensitivity and specificity obtained from the validation data. Thus by pooling or not-pooling the validation data across controls and cases, one can implement SIMEX under the nondifferential or differential misclassification assumptions respectively.

Results for all inferential schemes are reported in terms of mean-squared error of point estimators, and coverage and average width of nominal 95% interval estimators.

Results under the higher setting of exposure prevalence are given in Table 2 (MLE and Bayes) and Table 3 (SIMEX). Here nondifferential methods perform better when the data truly are nondifferentially misclassified (scenario 1). Differential methods are not as efficient under truly nondifferential misclassification, because sensitivity and specificity are estimated separately, and therefore via less data, for controls and cases. On the other hand, coverage of nondifferential methods deteriorates rapidly as the true misclassification mechanism becomes more differential. The performance of Bayes and ML methods is quite comparable, with the former having somewhat smaller MSE when differential misclassification is correctly assumed. Note also that in comparing Table 3 to Table 2, even the empirically better choice of extrapolation function (quadratic rather than loglinear) in SIMEX gives much larger MSE than the corresponding Bayes or ML procedure, particulary for the differential misclassification scenarios. This is not necessarily surprising. While the SIMEX approach is intuitively appealing, it does not carry the large-sample efficiency guarantees that come with likelihood-based procedures.

Results for ML and Bayes procedures in the lower exposure prevalence setting are given in Table 4. The combination of rare exposure, relatively high sensitivity, and relatively small validation sample size implies that for some generated datasets no subject is truly exposed to the risk factor in the case or control group, i.e. $a_{i1} = 0$ and $a_{i2} = 0$. This leads to nonsensical ML results, either analytically in the differential case or numerically in the nondifferential case. Thus the ML-DF results in the table are based on only those datasets without such empty cells in the validation data. Results are not given in the table for the ML-NDF method, since the numerical optimization can fail (or give a dubious result) for datasets with ‘near-empty’ cells in the validation data, but it is unclear how to definitively divide the generated datasets into those which do and don’t suffer from this problem. In general, the problem of nearly or exactly empty cells does limit the utility of ML procedures, particularly given that rare exposures and small validation sample-sizes are common in epidemiological settings. In contrast, the performance of the Bayesian procedures evidenced in Table 4 seems quite reasonable, with dramatic MSE reductions for the Bayes-DF inferences compared to ML-DF. The smoothing which results from combining prior distributions on sensitivity and specificity with empty or near-empty validation-data cells appears to yield much more satisfactory inferences. Results for SIMEX estimators in the low exposure setting are not shown, but again the overall performance is worse than Bayes and ML procedures, and the use of “plugged-in” sensitivity and specificity estimates leads to the “empty-cell” concerns as with ML methods.

Results for the nearly nondifferential (NNDF) Bayesian analysis, in both low and high exposure prevalence settings, appear in Table 5. For the sake of comparison, results are also given here for a two-stage non-Bayesian procedure that we refer to as *test-then-estimate* (TTE). The first TTE step applied a likelihood ratio test to the validation data, with the null hypothesis that the binary exposure is nondifferentially misclassified. Then as the second step ML-DF or ML-NDF point and interval estimates are reported, depending on whether the null is rejected or not in the first step. Again for some datasets one or more empty validation cell a_{ij} results in zero- or one-valued estimate for sensitivity, specificity or exposure prevalence hence yields nonsensical likelihood ratios, so that TTE estimates and inferential results are reported for only a subset of the simulated datasets.

The number of discarded datasets is higher in Table 5 than in Tables 2 and 4, for more simulated datasets have one or more empty cell than those having empty (a_{i1}, a_{i2}) pair(s) simultaneously. To aid comparison, the Bayes-NNDF results are reported for both (i) the same subset of datasets as for TTE, and (ii) all datasets.

In terms of both point and interval estimate performance, Bayes-NNDF is seen to offer good performance across scenarios, particularly in relation to either Bayes-NDF or Bayes-DF applied in a "wrong" scenario. Bayes-NNDF is also seen to be moderately better than TTE (in terms of both MSE and coverage) in the high exposure prevalence setting, and very substantially better than TTE in the low prevalence setting.

4 Example: Maternal use of antibiotics during pregnancy and sudden infant death syndrome

We consider a case-control study on sudden infant death syndrome (SIDS) [14] to further illustrate how Bayes, ML and SIMEX adjustments for misclassification work in practice. During investigation of a potential impact of maternal use of antibiotics during pregnancy on the occurrence of SIDS, surrogate exposure X was obtained from an interview question (yes=1, no=0). Information on antibiotic use from medical records, taken to be the actual exposure status T, was extracted for a subset of study participants. The data are shown in Table 6. Ignoring misclassification, the $X - Y$ log odds ratio is estimated as 0.352 with 95% confidence interval (0.101, 0.603).

The same prior distributions used in the simulation studies of Section 3 are employed here for drawing Bayesian inferences. Study results after the various adjustments for misclassification are presented in Table 7. Point and interval estimates of log-OR via Bayes and ML methods are similar. Parameters are estimated with more slightly more certainty under the nondifferential assumption than under the differential assumption, which is consistent with simulation findings.

Note that a considerably stronger exposure-disease association is estimated under the nondifferential misclassification assumption than under the differential misclassification assumption, with 'significance' (i.e., interval estimate excluding zero) in the former case but not the latter. Moreover, the validation-data evidence concerning differentiality is equivocal (likelihood ratio test P-value of 0.096 for the null hypothesis of nondifferential misclassication). Therefore, the Bayes-NNDF analysis may be viewed as an appropriate compromise between the nondifferential and differential analyses, with a tempered point estimate (relative to NDF) but still significant interval estimate.

The behaviour of SIMEX estimates in this example requires some comment. The extrapolation of the estimated log odds-ratio as a function of the misclassification magnitude is depicted in Figure 1. In line with the Bayes and ML results, adding further misclassification pushes estimates toward the null in the nondifferential case but away from the null in the differential case. In the nondifferential case, both choices of extrapolation function appear to fit the simulated data well. Extrapolating back to the no misclassification setting, however, produces adjusted estimates which are much more extreme than those obtained by either Bayes or ML methods.

5 Example: HSV-2 and invasive cervical cancer

The second example describes a case-control study consisting of 732 subjects of cervical cancer and 1312 community or hospital controls with negative cervical cancer diagnosis [15]. Researchers were interested in assessing the impact of herpes simplex virus type 2 (HSV-2, a binary variable) in the development of invasive cervical cancer. The exposure status was detected by the western blot assay, which produced error-prone measurements.

A refined, more accurate procedure was performed on a randomly selected sample of study subjects (selected without regard to their disease status), in order to assess the misclassification rates. The data are displayed in Table 8. It is noticeable from the validation and main data that the exposure prevalence of HSV-2 is high in both cases and controls. Carroll et al. observed from the validation sample that the misclassification differs between cases and controls (Fisher's exact two-sided test implied a greater sensitivity for the cases, p=0.049), and proposed a pseudo-likelihood model to adjust for the differential measurement error [16].

Ignoring measurement error arising from the inaccurate western blot procedure, the naive log odds ratio is estimated as 0.453 (standard error = 0.093), with 95% confidence interval (0.271, 0.635), indicating HSV-2 is positively correlated with the occurrence of invasive cervical cancer. We conduct Bayesian adjustment under three misclassification situations (NDF, NNDF and DF), again using the prior distributions for *logit* transformed sensitivities and specificities described in Section 3. For $logit(r_i)$, a flat prior with large variance is used here to generate posterior inference (μ=-1.946, σ=100). Similar results are observed when same hyperparameters for $logit(r_i)$ stated in Section 3 are used (μ=-1.946, σ=0.993).

Table 9 presents results of the various analyses. For all three methods (Bayes, ML, SIMEX), moving from the nondifferential assumption to the differential assumption moves the point estimate of the exposure-disease association toward the null, and causes the left endpoint of the interval estimate to move from positive to negative, i.e., "significance" is lost. (More precisely, for SIMEX this occurs under the quadratic extrapolation but not under the loglinear extrapolation, with Figure 2 suggesting that the quadratic extrapolation is more appropriate.)

As Carroll, Gail, and Lubin [16] pointed out, there is moderate evidence to show measurement error is differential across cases and controls. Sensitivities estimated from validation data alone are 0.78 for cases and 0.5 for controls. Nevertheless, if both the complete and incomplete data are considered, a likelihood ratio test for the nondifferentiality of misclassification with 2 degrees of freedom, generates a p-value at 0.073, indicating lack of evidence to reject the null at 5% significance level. The same test based merely on the validation data reports a consistent result (p-value= 0.084). Hence, it seems more appropriate to interpret the differentiality of measurement as borderline. One advantage of Bayesian adjustment emerges in this context, as it can incorporate the "in-between" scenario of nearly nondifferential misclassification via an appropriate prior distribution. As expected, the NNDF analyis yields a posterior mean and SD falling in between those arising from the NDF and DF assumptions. The resulting interval estimate is wholly positive, providing evidence for a positive exposure-disease association without concern about imposing an overly-strong assumption of nondifferential misclassification.

As a final point, we note that the Bayesian parameter estimates are consistent with the results given by Skrondal and Rabe-Hesketh [6] for these data, using generalized latent variable modeling techniques.

6 Discussion

Mismeasurement of exposure is an issue of broad concern in epidemiological studies, and there is a substantial literature on adjusting inferences on exposure-disease relationships in light of such mismeasurement. Bayesian methods, likelihood methods, and SIMEX methods are three general tools for implementing such adjustments. At least in the context of misclassified binary exposure, this paper has illustrated several positive attributes of the Bayesian approach. First, Bayesian methods can provide more reasonable and stable inferences when the resulting data are sparse, which is of particular relevance to small validation datasets in rare exposure contexts. Second, the infusion of prior information offered by the Bayesian approach can be used to good effect. Rather than committing to nondifferential or 'fully' differential assumptions concerning the exposure misclassification, a prior can be constructed to represent a 'nearly nondifferential' assumption. That is, the analyst can assert that substantial deviations from nondifferentiality are unlikely. This would seem to be a particularly useful device when the data themselves do not clearly support or refute nondifferentiality, as occured in both our real-data examples.

References

[1] B. A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.

[2] R. J. Marshall. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43:941–947, 1990.

[3] R. H. Lyles. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*, 58:1034–1037, 2002.

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models*, volume 105 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, Boca Raton, second edition, 2006.

[5] H. Küchenhoff, S. M. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, 62:85–96, 2006.

[6] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: multilevel,*

longitudinal, and structural equation models. Chapman & Hall/CRC, Boca Raton, 2004.

[7] P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, volume 13 of *Interdisciplinary Statistics.* Chapman & Hall/CRC, Boca Raton, 2004.

[8] G. J. Prescott and P. H Garthwaite. A simple bayesian analysis of misclassified binary data with a validation substudy. *Biometrics*, 58:454–458, 2002.

[9] P. Mclnturff, W. O. Johnson, D. Cowling, and I. A. Gardner. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine*, 23:1095–1109, 2004.

[10] M. Ladouceur, E. Rahme, C. A. Pineau, and J. Lawrence. Robustness of prevalence estimates derived from misclassified data from administrative aatabases. *Biometrics*, 63:272–279, 2007.

[11] S. Greenland. Sensitivity analysis, monte carlo risk analysis, and bayesian uncertainty assessment. *Risk Analysis*, 21:579–583, 2001.

[12] J. R. Cook and L. A. Stefansk. Simulation-extrapolation estimation in parametric measurement error eodels. *Journal of the American Statistical Association*, 89:1314–1328, 1994.

[13] W. Lederer and H. Küchenhoff. Simex: simex- and mcsimex- algorithm for measurement error models (version1.2) [software]. Available from http://cran.r-project.org/web/packages/simex/index.html.

[14] J. F. Kraus, S. Greenland, and M. Bulterys. Risk factors for sudden infant death syndrome in the us collaborative penrinatal project. *International Journal of Epidemiology*, 18:113–120, 1989.

[15] A. Hildesheim, V. Mann, L. A. Brinton, M. Szklo, W. C. Reeves, and W. E. Rawls. Herpes simplex virus type 2: a possible interaction with human papillomavirus type 16/18 in the development of invasive cervical cancer. *international Journal of Cancer*, 49:335–340, 1991.

[16] R. J. Carroll, M. H. Gail, and J. H. Lubin. Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88:185–199, 1993.

Table 1: Validation data and main data

	Validation Data				*Main Data*			
	Y=1		Y=0		Y=1		Y=0	
T	X=1	X=0	X=1	X=0	X=1	X=0	X=1	X=0
T=1	a_{11}	a_{12}	a_{01}	a_{02}	b_{11}	b_{12}	b_{01}	b_{02}
T=0	a_{13}	a_{14}	a_{03}	a_{04}	b_{13}	b_{14}	b_{03}	b_{04}
N	$a_{11} + a_{13}$	$a_{12} + a_{14}$	$a_{01} + a_{03}$	$a_{02} + a_{04}$	a_{15}	a_{16}	a_{05}	a_{06}

Table 2: Comparative performance of MLE and Bayesian models on simulated data (N_{rep}=400) given high exposure prevalences

		ML-NDF	ML-DF	Bayes-NDF	Bayes-DF
Scenario 1	MSE	0.035	0.071	0.034	0.054
	Coverage	0.968	0.950	0.963	0.963
	Width	0.768	1.066	0.779	1.018
Scenario 2	MSE	0.058	0.095	0.055	0.076
	Coverage	0.925	0.94	0.915	0.948
	Width	0.823	1.112	0.823	1.054
Scenario 3	MSE	0.092	0.095	0.083	0.076
	Coverage	0.863	0.945	0.855	0.960
	Width	0.873	1.145	0.860	1.078
Scenario 4	MSE	0.138	0.083	0.120	0.067
	Coverage	0.798	0.950	0.790	0.958
	Width	0.918	1.170	0.896	1.102

Table 3: Performance of SIMEX on simulated data (N_{rep}=400) given high exposure prevalences

		SIMEX-NDF				SIMEX-DF			
		Quadratic extrapolation		Loglinear extrapolation		Quadratic extrapolation		Loglinear extrapolation	
		Asymp. var.	Jackknife var.	Asymp. var.	Jackknife var.	Asymp. var.	Jackknife var.	Asymp. var.	Jackknife var.
Scenario 1	MSE	0.036	0.036	0.040	0.040	0.113	0.113	0.738	0.738
	Coverage	0.960	0.940	0.960	0.938	0.757	0.728	0.688	0.755
	Width	0.784	0.729	0.800	0.729	0.787	0.730	0.843	0.730
Scenario 2	MSE	0.053	0.053	0.068	0.068	0.138	0.138	0.649	0.649
	Coverage	0.935	0.893	0.900	0.860	0.735	0.683	0.725	0.790
	Width	0.833	0.756	0.871	0.756	0.828	0.752	0.945	0.752
Scenario 3	MSE	0.107	0.107	0.165	0.165	0.147	0.147	1.424	1.424
	Coverage	0.840	0.768	0.763	0.655	0.753	0.695	0.810	0.845
	Width	0.887	0.786	0.956	0.786	0.879	0.777	0.848	0.777
Scenario 4	MSE	0.202	0.202	0.366	0.366	0.161	0.161	0.503	0.503
	Coverage	0.663	0.555	0.555	0.395	0.750	0.678	0.830	0.828
	Width	0.939	0.809	1.058	0.809	0.933	0.792	0.767	0.792

Table 4: Comparative performance of MLE(N_{rep}<400) and Bayesian (N_{rep}=400) models on simulated data given low exposure prevalences. For the ML-DF procedure, the number in parentheses indicates the number of datasets excluded due to having zero truly exposed subject in a group.

		ML-DF	Bayes-NDF	Bayes-DF
Scenario 1	MSE	0.535 (-9)	0.182	0.228
	Coverage	0.928	0.960	0.978
	Width	2.418	1.849	2.156
Scenario 2	MSE	0.501 (-17)	0.258	0.217
	Coverage	0.945	0.920	0.980
	Width	2.416	1.865	2.184
Scenario 3	MSE	0.512 (-18)	0.687	0.206
	Coverage	0.940	0.648	0.985
	Width	2.470	1.861	2.218
Scenario 4	MSE	0.499 (-16)	1.229	0.210
	Coverage	0.935	0.308	0.980
	Width	2.4840	1.836	2.226

Table 5: Comparative performance of MLE TTE and Bayesian NNDF on the subsets of simulated data

		High prevalences				Low prevalences			
		m_h	ML-TTE (N_{rep}=m_h)	Bayes-NNDF (N_{rep}=m_h)	Bayes-NNDF (N_{rep}=400)	m_l	ML-TTE (N_{rep}=m_l)	Bayes-NNDF (N_{rep}=m_l)	Bayes-NNDF (N_{rep}=400)
Scenario 1	MSE	393	0.039	0.036	0.035	104	4.223	0.162	0.183
	Coverage		0.962	0.967	0.968		0.981	0.981	0.978
	Width		0.782	0.834	0.834		2.065	1.837	1.958
Scenario 2	MSE	394	0.065	0.054	0.055	134	0.338	0.107	0.184
	Coverage		0.924	0.942	0.938		0.985	1	0.983
	Width		0.855	0.882	0.883		2.181	1.901	2.006
Scenario 3	MSE	397	0.090	0.069	0.069	135	4.014	0.224	0.316
	Coverage		0.894	0.909	0.908		0.716	0.970	0.940
	Width		0.944	0.924	0.924		2.295	1.978	2.069
Scenario 4	MSE	391	0.106	0.083	0.083	146	0.908	0.309	0.451
	Coverage		0.877	0.918	0.920		0.678	0.918	0.863
	Width		1.050	0.969	0.968		2.440	2.007	2.109

Table 6: Validation study and main study of SIDS

	Validation data				*Main data*		
	Y=1		Y=0				
T	X=1	X=0	X=1	X=0	Y	X=1	X=0
T=1	29	17	21	16	Y=1	122	442
T=0	22	143	12	168	Y=0	101	479

Table 7: $log\widehat{OR}$, SD and 95% intervals for $logOR$ in SIDS study

	$log(\widehat{OR})$	SD	95% intervals
Bayes-NDF	0.395	0.187	(0.034, 0.763)
Bayes-NNDF	0.303	0.198	(-0.079, 0.698)
Bayes-DF	0.211	0.215	(-0.219, 0.630)
ML-NDF	0.398	0.191	(0.024, 0.772)
ML-DF	0.193	0.221	(-0.241, 0.626)
SIMEX-NDF			
quadratic, asymptotic	0.663	0.227	(0.219, 1.108)
quadratic, Jackknife		0.179	(0.313, 1.014)
loglinear, asymptotic	0.725	0.257	(0.220, 1.229)
loglinear, Jackknife		0.179	(0.374, 1.075)
SIMEX-DF			
quadratic, asymptotic	-0.010	0.221	(-0.443, 0.424)
quadratic, Jackknife		0.209	(-0.418, 0.399)
loglinear, asymptotic	0.276	0.110	(0.060, 0.492)
loglinear, Jackknife		0.209	(-0.133, 0.685)

Table 8: Validation data and main data for cervical cancer study

	Validation data				*Main data*		
	Y=1		Y=0				
T	X=1	X=0	X=1	X=0	Y	X=1	X=0
T=1	18	5	16	16	Y=1	375	318
T=0	3	13	11	33	Y=0	535	701

Table 9: $log\widehat{OR}$, SD and 95% intervals for $logOR$ in cervical cancer study based on a flat prior for logit(r_i)

	$log(\widehat{OR})$	SD	95% interval
Bayes-NDF	0.921	0.223	(0.520, 1.404)
Bayes- NNDF	0.809	0.266	(0.320, 1.356)
Bayes-DF	0.583	0.324	(-0.033, 1.262)
ML-NDF	0.958	0.237	(0.494, 1.422)
ML-DF	0.608	0.350	(-0.079, 1.295)
SIMEX-NDF			
quadratic, asymptotic	0.903	0.184	(0.542, 1.264)
quadratic, Jackknife		0.133	(0.643, 1.164)
loglinear, asymptotic	1.198	0.239	(0.730, 1.667)
loglinear, Jackknife		0.133	(0.938, 1.459)
SIMEX-DF			
quadratic, asymptotic	0.146	0.172	(-0.191, 0.482)
quadratic, Jackknife		0.137	(-0.123, 0.414)
loglinear, asymptotic	0.427	0.100	(0.231, 0.623)
loglinear, Jackknife		0.137	(0.158, 0.696)

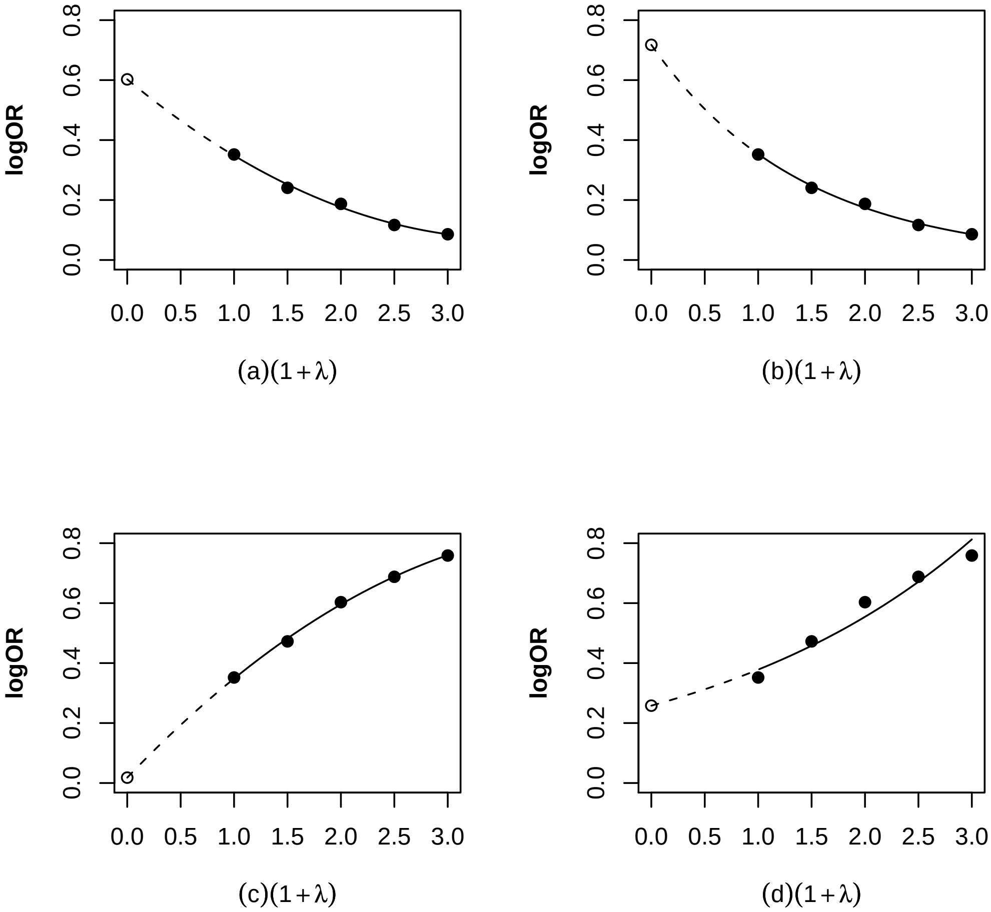

Figure 1: *Plots of the estimated logOR as a function of misclassification size λ in SIDS study. The upper-left panel is based on a quadratic extrapolation subject to NDF MC-SIMEX. The upper-right panel is based on a loglinear extrapolation subject to NDF MC-SIMEX. The lower-left panel is based on a quadratic extrapolation subject to DF MC-SIMEX. The lower-right panel is based on a loglinear extrapolation subject to DF MC-SIMEX.*

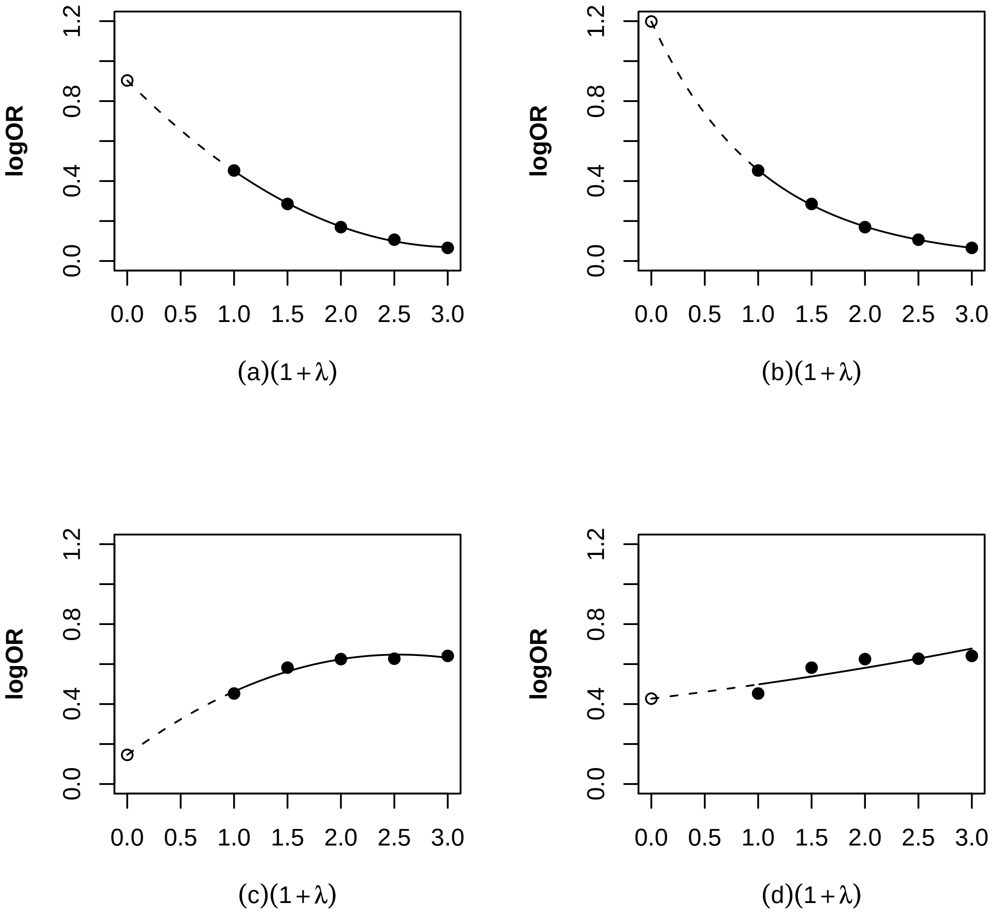

Figure 2: *Plots of the estimated logOR as a function of misclassification size λ in cervical cancer study. The upper-left panel is based on a quadratic extrapolation subject to NDF MC-SIMEX. The upper-right panel is based on a loglinear extrapolation subject to NDF MC-SIMEX. The lower-left panel is based on a quadratic extrapolation subject to DF MC-SIMEX. The lower-right panel is based on a loglinear extrapolation subject to DF MC-SIMEX.*