Published in final edited form as:
Stat Med. 2010 September 30; 29(22): 2297–2309. doi:10.1002/sim.3971.

Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting

Robert H. Lyles[*,†] and **Ji Lin**
Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322, U.S.A.

Abstract

The potential for bias due to misclassification error in regression analysis is well understood by statisticians and epidemiologists. Assuming little or no available data for estimating misclassification probabilities, investigators sometimes seek to gauge the sensitivity of an estimated effect to variations in the assumed values of those probabilities. We present an intuitive and flexible approach to such a sensitivity analysis, assuming an underlying logistic regression model. For outcome misclassification, we argue that a likelihood-based analysis is the cleanest and the most preferable approach. In the case of covariate misclassification, we combine observed data on the outcome, error-prone binary covariate of interest, and other covariates measured without error, together with investigator-supplied values for sensitivity and specificity parameters, to produce corresponding positive and negative predictive values. These values serve as estimated weights to be used in fitting the model of interest to an appropriately defined expanded data set using standard statistical software. Jackknifing provides a convenient tool for incorporating uncertainty in the estimated weights into valid standard errors to accompany log odds ratio estimates obtained from the sensitivity analysis. Examples illustrate the flexibility of this unified strategy, and simulations suggest that it performs well relative to a maximum likelihood approach carried out via numerical optimization.

Keywords

bias; errors-in-variables; odds ratio; regression

1. Introduction

As a special case of covariate measurement error [1], the misclassification of categorical predictor variables in regression analysis is known to pose a potentially serious threat to the validity of parameter estimates and statistical inferences. Errors in measuring binary outcome variables are an equally real possibility in applied research. As such, misclassification has long been a topic of interest to statisticians and epidemiologists, beginning with classic papers characterizing its effects upon the estimation of odds ratios based on tabular data [2--4]. These references provided groundwork for methods incorporating validation [5--7] or reproducibility data (e.g. [8, 9]) into statistical corrections in order to obtain valid point and interval estimates.

[*] Correspondence to: Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322, U.S.A. . [†] rlyles@sph.emory.edu .

Often, however, investigators may suspect the potential for misclassification but have no auxiliary data available for estimating sensitivity and specificity parameters upon which one would base a correction. In such cases researchers might turn to sensitivity analysis, in which these parameters would generally be assumed to be known and then a value for the desired measure of effect (usually an odds ratio) is inferred based on those assumptions. By varying the assumed values of the misclassification parameters, one can get a feel for the feasible range of the association parameter of interest and for the extent to which its estimated value based on the error-prone covariate data may be misleading.

Examples of sensitivity analysis in the misclassification context are prevalent in the recent epidemiologic and statistical literature [10--14]. We agree with these previous authors about the utility of such analyses, and the importance of making tools for conducting them available for statistical and epidemiologic research. Most prior work has focused on standard tabular data, with a binary risk factor subject to misclassification. Fox *et al*. [11] offered an extension by providing a computer macro to apply sensitivity analysis toward the odds ratio associated with a misclassified binary variable, adjusted for other covariates via logistic regression. Here, we seek to further extend the focus within the logistic regression setting while enhancing the conceptual and computational accessibility of such sensitivity analyses.

Our approach is to treat the problems of outcome and covariate misclassification separately. In the former case, we find a likelihood-based analysis to be straightforward and consistent with prior literature [15]. When a predictor variable is misclassified, however, the choice of an ideal approach is less clear-cut. In this case our goal is to propose a unified methodological framework that allows one to account for implied restrictions on parameters, and to generalize sensitivity analyses into new and important directions. While some programming is required, our recommended approach is less simulation-driven than prior proposals [10, 11] and we provide sufficient detail to make it readily programmable by potential users. Estimated odds ratios reflecting the assumed misclassification rates are obtained using standard software for logistic regression, with appropriately defined weights. The result is a complete set of estimated regression parameters corresponding to the desired logistic regression model, with jackknife-based standard errors recommended to properly account for variability in the observed data. The weighting approach provides welcome flexibility and generalizations, and we draw comparisons between it and direct maximum likelihood (ML) as an alternative.

A note about the approaches to be described below is that their applicability is somewhat dependent upon study design considerations. In general, the methods aimed at adjusting for covariate misclassification are most readily applicable in cross-sectional or retrospective (case-control) study designs, as opposed to studies in which the distribution of exposure is controlled by the investigator. Similarly, methods aimed at outcome misclassification tend to be most suitable under cross-sectional or prospective sampling schemes.

2. Methods

2.1. The no-covariate case: basic sensitivity analysis

To begin, consider estimating a crude odds ratio corresponding to a potentially misclassified binary risk factor. Table I revisits data from an often-cited case-control study of the relationship between maternal antibiotic use during pregnancy and the odds of sudden infant death syndrome (SIDS) [5--7, 16]. The observed exposure data reflect self-reported antibiotic use (Z), whereas the aim is to associate the true binary variable of antibiotic use (X) with SIDS status (Y).

The observed odds ratio (OR) associated with Table I is 1.309, corresponding to an estimated ln(OR) (std. error) of 0.269 (0.150). A sensitivity analysis to examine potential effects of misclassification would be based on one or more sets of assumptions about the underlying sensitivity (SE) and specificity (SP) parameters. These consist of SE=Pr(Z =1|X =1) and SP=Pr(Z =0|X =0) or, for differential misclassification [17], SE_y=Pr(Z =1|X =1, Y = y) and SP_y=Pr(Z =0|X =0, Y = y)(y=0, 1).

In this simple 2×2 case-control setting, the most straightforward approach to sensitivity analysis may be to base it directly on writing familiar ‘matrix-method’-type identities in terms of probabilities [3, 18]. In particular, it is easily shown that

$$\pi_y=\frac{\pi_y^*+\text{SP}_y-1}{\text{SE}_y+\text{SP}_y-1}, \tag{1}$$

where π_y =Pr(X =1|Y = y) and $\pi_y^*=\Pr(Z=1|Y=y)$ (y =0,1). Given estimates of π_y^* from the observed data, one may supply values for SE_y and SP_y and then calculate the resulting true exposure probabilities via (1). The estimate $(\widehat{\text{OR}})$ follows as $\pi_1(1-\pi_0)[\pi_0(1-\pi_1)]^{-1}$, where typically we would examine several (SE_y, SP_y) combinations to obtain a range of estimated ORs. The delta method may readily be used to estimate the variance of each ln(OR) estimate, accounting for uncertainty in the estimates $\widehat{\pi}_1^*$ and $\widehat{\pi}_0^*$ based on the observed data.

Equation (1) is also useful in that it implies two important restrictions that should be considered when conducting the analysis:

$$\pi_y^*<\text{SE}_y \quad \text{and} \quad \pi_y^*>1-\text{SP}_y. \tag{2}$$

Choosing values of SE_y and SP_y that are incompatible with these restrictions (upon replacing the π_y^*’s by their estimates) will generally lead to numerical issues with the sensitivity analysis.

As an alternative procedure for determining the sensitivity analysis-based log(OR) estimate and its standard error, one could also adopt an ML approach. For instance, assume the underlying logistic regression model of interest:

$$\text{logit}\,[\Pr(Y=1|X=x)]=\alpha+\beta x \quad (x=0,1), \tag{3}$$

where primary interest is in the log(OR) parameter (β). Observed-data likelihood contributions corresponding to investigator-supplied SE_y and SP_y values are expressible in a traditional manner applicable to covariate measurement error settings [1], with a slight adjustment to account for potentially differential misclassification:

$$\Pr(Y=y,Z=z)=\sum_{x=0}^{1}\Pr(Z=z|X=x,Y=y)\Pr(Y=y|X=x)\Pr(X=x). \tag{4}$$

Note that the first term on the right side of (4) is determined by the assumed (SE_y, SP_y) values, whereas the second term follows from model (3) and the last is an estimable nuisance parameter. Numerically maximizing the log-likelihood determined by the contributions in (4) provides $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)$ with its standard error estimable by approximating the observed information matrix.

In the following section, we outline a third approach to sensitivity analysis for covariate misclassification that facilitates implementation and generalization, while making an intriguing alternative to the direct ML approach based on equation (4).

2.2. The no-covariate case: predictive value weighting

While basing sensitivity analysis on the matrix-method identity in (1) may offer the most direct route in the no-covariate setting, several authors [10, 11] have proposed methods based on positive and negative predictive values. The basic idea is to 'reconstruct' data that might have been observed under no misclassification, by using the observed data (e.g. Table I) together with assumptions about (SE_y, SP_y) to arrive at estimates of $\mathrm{PPV}_y = \Pr(X=1|Z=1, Y=y)$ and $\mathrm{NPV}_y = \Pr(X=0|Z=0, Y=y)$. Once these predictive value estimates based on the observed data and the assumed (SE, SP) values are in hand, we note that the following identity analogous to (1) may be applied to execute sensitivity analysis:

$$\pi_y=\mathrm{PPV}_y\pi_y^*+\left(1-\mathrm{NPV}_y\right)\left(1-\pi_y^*\right)\quad (y=0,1) \tag{5}$$

This identity underlies the 'inverse' matrix method [6], which proves particularly useful when adjusting for differential misclassification in the presence of internal validation data [18, 19].

The use of (5) for sensitivity analysis is less direct than using (1), because of an additional step needed to obtain the appropriate PPV_y and NPV_y values by combining the assumed (SE_y, SP_y) combinations with the estimates of π_y^* ($y=0,1$). One can show that PPV_y and NPV_y are found as the solution to two linear equations, represented compactly as follows:

$$\mathbf{P}_y=\mathbf{A}_y^{-1}\mathbf{J}, \tag{6}$$

where

$$\mathbf{P}_y=\begin{pmatrix}\mathrm{PPV}_y\\ \mathrm{NPV}_y\end{pmatrix},\quad \mathbf{A}_y=\begin{pmatrix}\left(\mathrm{SE}_y-1\right)\pi_y^*\left[\mathrm{SE}_y\left(\pi_y^*-1\right)\right]^{-1} & 1\\ 1 & \left(\mathrm{SP}_y-1\right)\left(\pi_y^*-1\right)\left(\mathrm{SP}_y\pi_y^*\right)^{-1}\end{pmatrix}\quad\text{and}\quad \mathbf{J}=\begin{pmatrix}1\\ 1\end{pmatrix}$$

and we replace the π_y^*'s by observed data-driven estimates. Use of this result in conjunction with (5) would also complicate standard error calculations relative to using (1). However, the estimated predictive values in (6) are the key to the approach advocated henceforth, in which we use them as weights for fitting a desired logistic regression model while adjusting for covariates.

To outline the proposed approach in the no-covariate case, consider viewing each (Z,Y) record in Table I as reflecting two possible (X,Y) records (one with $x=0$, the other with $x=1$). Thus, the four possible (Z,Y) pairs yield a total of eight (X,Z,Y) combinations, each of which we weight appropriately as indicated in Table II. The fourth column in Table II defines the predictive value weights that are central to our approach, whereas the two right-most columns provide the estimated values of these weights corresponding to the observed data in Table I for the non-differential and differential example scenarios detailed in Section 3.1.

The appeal of this approach is that the desired model associating *Y* with the true binary exposure variable (*X*) may now be fit directly to the (*y*, *x*) data in Table II. Formally, we maximize the weighted log-likelihood given by

$$l(\theta)=\sum_{z=0}^{1}\sum_{y=0}^{1}\left\{\sum_{i=1}^{n_{zy}}\sum_{x=0}^{1}w_{xzy}l_{xy}(\theta)\right\}, \tag{7}$$

where the n_{zy}'s represent the observed-data cell counts where $Z = z$ and $Y = y$ ($z, y = 0, 1$) and $l_{xy}(\boldsymbol{\theta})$ is the log-likelihood contribution for a (*y*, *x*) pair. In practice, this simply means that we fit the logistic regression model (3) to the (*y*, *x*) data in Table II, applying the weights (w_{xzy}). This can be accomplished via standard software, e.g. by direct application of the SAS LOGISTIC procedure with a 'WEIGHT' statement [20] to the expanded data set. This straightforward one-step weighting process removes any need for simulation (e.g. [11]). Note that we fit the logistic model to a total of twice the number of observations as originally obtained (i.e. 2288 for the Table I example), since each observation is replicated to allow either $x = 0$ or $x = 1$.

To obtain a standard error accounting for the uncertainty in the observed-data estimates of π_y^* (y=0, 1) used to compute the weights, we suggest a standard bootstrap [21] or jackknife [21, 22] procedure. We use the jackknife estimator henceforth, as we find it much less susceptible to numerical problems. Specifically, the potential for conflicts with the necessary restrictions in (2) is very small when employing the jackknife but can be substantial under bootstrap sampling. For each 'leave one out' sample from the original observed data, we recalculate the weights via (6) and re-fit the weighted [summationtext] [summationtext] logistic regression in (7) using standard software. The jackknife standard error is calculated based on the $n=\sum_{z=0}^{1}\sum_{y=0}^{1}n_{zy}$ resulting $\boldsymbol{\theta}$ estimates.

2.3. Outcome misclassification in the no-covariate case

When the outcome (*Y*) rather than the exposure (*X*) is potentially misclassified, the problem has been more broadly discussed in the literature [15, 23, 24]. Magder and Hughes [15] facilitated EM algorithm-based ML analysis via weighting of observations in a manner somewhat akin to our proposal for risk factor misclassification in the previous section. We favor a direct ML approach for outcome misclassification, given the straightforward manner in which the appropriate likelihood can be specified and numerically maximized. Assume that we desire to fit model (3), but that an error-prone variable Y^* is observed in place of the true binary outcome *Y*. The sensitivity and specificity of Y^* are given by $\mathrm{SE}_x = \Pr(Y^* = 1|Y = 1, X = x)$ and $\mathrm{SP}_x = \Pr(Y^* = 0|Y = 0, X = x)$ ($x = 0, 1$). Observations contribute to the likelihood as follows:

$$\Pr(Y^*=y^*|X=x)=\sum_{y=0}^{1}\Pr(Y^*=y^*|Y=y, X=x)\Pr(Y=y|X=x), \tag{8}$$

with the first term given by the assumed (SE_x, SP_x) values and the second by the desired model (3). This function is readily handled using standard optimization routines such as those available in SAS IML [25], Splus [26], and R [27]. Standard errors for all parameters in model (3) are also readily obtained using such software via close numerical approximations to the observed information matrix. Note that when specifying SE_x and SP_x, restrictions identical to those in (2) should be observed, except with ‘*x*’ replacing ‘*y*’ in the subscripts.

2.4. Adjusting for covariates: predictive value weighting and ML

We now turn to the more common setting in which interest lies in one or more covariate-adjusted odds ratios, normally estimable by the following multiple logistic regression model:

$$\mathrm{logit}\left[\Pr\left(Y=1|X,\mathbf{C}\right)\right]=\alpha+\beta x+\sum_{j=1}^{J}\gamma_j c_j \quad (x=0,1)\,. \tag{9}$$

As before, we assume that *X* is subject to misclassification and that one only has access to the error-prone binary variable *Z*. We assume that the covariates $C_j (j=1, \ldots, J)$ are observed without error.

Although one might initially contemplate covariate-adjusted sensitivity analysis using a generalization of equation (1), there are some fundamental drawbacks to such a strategy. A more promising approach is to generalize the likelihood in equation (4). In this case, the investigator specifies the parameters $\mathrm{SE}_{\mathrm{y}\mathbf{c}}=\Pr(Z=1|X=1, Y=y, \mathbf{C}=\mathbf{c})$ and $\mathrm{SP}_{\mathrm{y}\mathbf{c}}=\Pr(Z=0|X=0, Y=y, \mathbf{C}=\mathbf{c})$ $(x, y, z=0, 1)$. Observed-data likelihood contributions corresponding to these specified values are given by $\Pr(Y=y, Z=z|\mathbf{C}=\mathbf{c})$, which may be expressed as follows:

$$\sum_{x=0}^{1}\Pr\left(Z=z|X=x,Y=y,\mathbf{C}=\mathbf{c}\right)\Pr\left(Y=y|X=x,\mathbf{C}=\mathbf{c}\right)\Pr\left(X=x|\mathbf{C}=\mathbf{c}\right)\,. \tag{10}$$

As before, the first term in (10) reflects the assumed $(\mathrm{SE}_{y\mathbf{c}}, \mathrm{SP}_{y\mathbf{c}})$ values, whereas the second term reflects the desired model (i.e. (9)). The last term might be modeled, e.g., via a second logistic regression model of *X* on **C**.

Alternatively, the predictive value weighting approach introduced in Section 2.2 remains conceptually straightforward and accessible in the covariate-adjusted case based on the following multivariable extensions of (5) and (6):

$$\pi_{y\mathbf{c}}=\mathrm{PPV}_{y\mathbf{c}}\pi^*_{y\mathbf{c}}+\left(1-\mathrm{NPV}_{y\mathbf{c}}\right)\left(1-\pi^*_{y\mathbf{c}}\right) \quad (y=0,1) \tag{11}$$

and

$$\mathbf{P}_{y\mathbf{c}}=A^{-1}_{y\mathbf{c}}\mathbf{J}, \tag{12}$$

where

$$\mathbf{P}_{y\mathbf{c}}=\begin{pmatrix}\mathrm{PPV}_{y\mathbf{c}}\\ \mathrm{NPV}_{y\mathbf{c}}\end{pmatrix},\quad \mathbf{A}_{y\mathbf{c}}=\begin{pmatrix}\left(\mathrm{SE}_{y\mathbf{c}}-1\right)\pi^*_{y\mathbf{c}}\left[\mathrm{SE}_{y\mathbf{c}}\left(\pi^*_{y\mathbf{c}}-1\right)\right]^{-1} & 1\\ 1 & \left(\mathrm{SP}_{y\mathbf{c}}-1\right)\left(\pi^*_{y\mathbf{c}}-1\right)\left(\mathrm{SP}_{y\mathbf{c}}\pi^*_{y\mathbf{c}}\right)^{-1}\end{pmatrix},\quad \mathbf{J}=\begin{pmatrix}1\\ 1\end{pmatrix},$$

and $\pi^*_{y\mathbf{c}}=\Pr(Z=1|\mathbf{C}=\mathbf{c}, Y=y)$. Here, the predictive value parameter definitions have been generalized in the expected manner, i.e. $\mathrm{PPV}_{y\mathbf{c}} = \Pr(X=1|Z=1, Y=y, \mathbf{C}=\mathbf{c})$ and $\mathrm{NPV}_{y\mathbf{c}} = \Pr(X=0|Z=0, Y=y, \mathbf{C}=\mathbf{c})$. With the $(\mathrm{SE}_{y\mathbf{c}}, \mathrm{SP}_{y\mathbf{c}})$ parameters specified by the investigator, the method proceeds exactly as before. We propose estimating the $\pi^*_{y\mathbf{c}}$'s via logistic regression, with attention to model fitting. For example, the assessment of potential interactions between Y and the C_j's may validate the model

$$\text{logit}\left(\pi^*_{y\mathbf{c}}\right)=\text{logit}\left[\Pr\left(Z=1|Y,\mathbf{C}\right)\right]=\eta+\omega y+\sum_{j=1}^{J}\psi_j c_j. \tag{13}$$

Otherwise, model (13) may be enriched with product terms to account for interactions present in the observed data, in the interest of valid $\pi^*_{y\mathbf{c}}$ estimates for use in (12). Note that predictive value weighting requires a model for $Z|(Y, \mathbf{C})$, whereas ML via (10) requires a model for $X|\mathbf{C}$.

We emphasize again the need for caution with respect to restrictions implied by the extension to equation (1) to incorporate covariates. Analogous to (2), it follows that

$$\pi^*_{y\mathbf{c}}<\mathrm{SE}_{y\mathbf{c}} \quad \text{and} \quad \pi^*_{y\mathbf{c}}>1-\mathrm{SP}_{y\mathbf{c}}, \tag{14}$$

for all $(y, \mathbf{c})$ combinations. Investigator-supplied values of $\mathrm{SE}_{y\mathbf{c}}$ or $\mathrm{SP}_{y\mathbf{c}}$ that fail to honor these restrictions (when replacing the $\pi^*_{y\mathbf{c}}$'s by their estimates) can contribute to numerical problems with ML based on (10), and produce predictive value weights via (12) that are outside the (0, 1) range. Thus, it is reasonable to allow the observed data to inform one's (SE, SP) choices, as suggested in (14). When all predictor variables are categorical in nature, it is straightforward to incorporate these restrictions into the selection of $\mathrm{SE}_{y\mathbf{c}}$ or $\mathrm{SP}_{y\mathbf{c}}$. With one or more continuous predictor (C) variables, we suggest the use of the following adjustments for this purpose:

$$\mathrm{SE}_{y\mathbf{c}}=\max\left(\mathrm{SE}^0_{y\mathbf{c}}, \widehat{\pi}^*_{y\mathbf{c}}\right) \quad \text{and} \quad \mathrm{SP}_{y\mathbf{c}}=\max\left(\mathrm{SP}^0_{y\mathbf{c}}, 1-\widehat{\pi}^*_{y\mathbf{c}}\right), \tag{15}$$

where $\mathrm{SE}^0_{y\mathbf{c}}$ and $\mathrm{SP}^0_{y\mathbf{c}}$ are the *a priori* investigator-supplied values and $\widehat{\pi}^*_{y\mathbf{c}}$ is estimated, e.g., via the logistic regression of Z on $(Y, \mathbf{C})$ in (13). In the case of predictive value weighting, the adjustments in (15) simply equate to setting negative weights produced by the original $\left(\mathrm{SE}^0_{y\mathbf{c}}, \mathrm{SP}^0_{y\mathbf{c}}\right)$ values equal to 0, while setting weights in excess of 1 equal to 1.

2.5. Adjusting for covariates: outcome misclassification

For covariate-adjusted sensitivity analysis to outcome misclassification via model (9), we prefer a direct likelihood approach as in Section 2.3. The form for the likelihood contributions becomes

$$\Pr\left(Y^*=y^*|X=x, \mathbf{C}=\mathbf{c}\right)=\sum_{y=0}^{1}\Pr\left(Y^*=y^*|Y=y, X=x, \mathbf{C}=\mathbf{c}\right)\Pr\left(Y=y|X=x, \mathbf{C}=\mathbf{c}\right), \tag{16}$$

where the second term in the summation follows from (9) and the first term is determined by the user-specified values $\mathrm{SE}_{x\mathbf{c}}$ =Pr(Y^* =1|Y =1, X = x, **C=c**) and $\mathrm{SP}_{x\mathbf{c}}$ =Pr(Y^* =0|Y =0, X = x, **C=c**) (x=0, 1). Restrictions as in (14) apply, except again with ‘x’ replacing ‘y’ in the subscripts.

2.6. Extensions

An advantage of the predictive value weighting approach is that extensions to more complex scenarios are conceptually straightforward. For example, suppose one was interested in sensitivity analysis in a situation where two binary predictors (X_1 and X_2) are subject to misclassification via corresponding observed surrogates Z_1 and Z_2. The appropriate expanded data set would consist of four (y, x_1, x_2, **c**) records in place of each observed (y, z_1, z_2, **c**) record, with each observation weight of the form Pr($X_1 = x_1$, $X_2 = x_2$|$Z_1 = z_1$, $Z_2 = z_2$, $Y = y$, **C=c**). Though more challenging than in the case of a single misclassified predictor, the calculation of these weights in analogous fashion could be facilitated to the extent that the investigator is willing to make simplifying assumptions about the misclassification processes (e.g. independent and/or non-differential misclassification of X_1 and X_2). Similarly, suppose we were concerned with misclassification of the outcome (Y) in addition to one binary predictor (X). In that case, each observed (y^*, z, **c**) record could be replaced by four (y, x, **c**) records, with weights of the form P($Y = y$, $X = x$|$Y^* = y^*$, $Z = z$, **C=c**). Again, estimation of these weights might be facilitated by certain plausible simplifying assumptions. The form of the weights suggests that this latter scenario may only be defensible under cross-sectional study designs.

3. Examples

3.1. The no-covariate case

For the data in Table I, we have $\widehat{\pi}_1^*=0.2163\ \ (0.0173)$ and $\widehat{\pi}_0^*=0.1741\ \ (0.0157)$. Suppose the investigator assumes non-differential misclassification, with SE=0.6 and SP=0.9. It follows from (1) that $\widehat{\pi}_1=0.2326\ \ (0.0346)$ and $\widehat{\pi}_0=0.1482\ \ (0.0314)$ so that $\widehat{\mathrm{OR}}=1.742$ and $\ln\left(\widehat{\mathrm{OR}}\right)=0.555\ \ (0.315)$, with the latter standard error obtained by the delta method. In contrast, suppose we had specified the following differential conditions: SE_1=0.6, SP_1=0.8, SE_0=0.7, SP_0=0.9. Using (1), we obtain $\widehat{\pi}_1=0.0408\ \ (0.0433)$ and $\widehat{\pi}_0=0.1235\ \ (0.0262)$ so that $\ln\left(\widehat{\mathrm{OR}}\right)=-1.199\ \ (1,133)$, with $\widehat{\mathrm{OR}}=0.302$. Note the dramatic shift in implications assuming differential misclassification.

Applying direct ML (see equation (4)) to the data in Table I yields $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=0.555\,(0.316)$ and $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=-1.199\,(1.135)$ under the non-differential and differential assumptions outlined above, respectively. Thus, results via the matrix-method identity in (1) with a delta method-based standard error are effectively identical to ML utilizing (4).

Application of the SAS LOGISTIC procedure with a ‘WEIGHT’ statement [20] to the expanded data set in Table II produces solutions identical to those of the matrix-method and likelihood-based approach: $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=0.555$ and $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=-1.199$ for the non-differential and differential cases, respectively. To estimate standard errors accounting for uncertainty in $\widehat{\pi}_1^*$ and $\widehat{\pi}_0^*$, we obtained 1144 $\ln\left(\widehat{\mathrm{OR}}\right)$ estimates using the proposed weighting method, each one calculated by leaving out a unique (z_i, y_i) pair. The jackknife standard error estimate [22] is 0.318 in the non-differential case and 1.170 in the differential case, closely matching the estimates obtained via the matrix-method and ML approaches.

Figure 1 displays ln(OR) estimates determined by predictive value weighting, with bars indicating ±1 jackknife standard error, for several (SE, SP) combinations characterizing non-differential exposure misclassification in the SIDS study data in Table I. Note that this more extensive sensitivity analysis is particularly informative in that it indicates very little change in the ln(OR) estimate as SE varies for a fixed SP (left panel of Figure 1), but dramatic changes as SP varies for a fixed SE (right panel). In exploring a greater variety of (SE, SP) scenarios in practice, one may wish to consider displaying the results via contour plots (e.g. [12]).

To illustrate sensitivity analysis for outcome misclassification, let us now pretend that sampling for the SIDS study had been done in a prospective or cross-sectional manner and that the outcome (*Y*) rather than the exposure (*X*) was error-prone. We then view the data in Table I as depicting *Y** vs. *X*. Numerically maximizing the likelihood based on equation (8) assuming non-differential error with SE=0.6 and SP=0.9 yields $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=0.982\ (0.727)$, so that $\widehat{\mathrm{OR}}=2.67$. Assuming instead a differential scenario with SE_1=0.6, SP_1=0.8, SE_0=0.7, and SP_0=0.9 gives $\widehat{\beta}=\ln\left(\widehat{\mathrm{OR}}\right)=1.33\ (0.735)$, so that $\widehat{\mathrm{OR}}=3.78$. In both of these cases, the natïve OR estimate 1.31 would be markedly attenuated.

3.2. Adjusting for a binary covariate

For the purpose of illustration, Table III shows the original SIDS data of Table I, stratified based on an artificial binary covariate *C*. Equation (9) gives the desired model, with *X* as the predictor of primary interest and *C* as the sole covariate. Estimates of $\pi^*_{y\mathbf{c}}$'s were obtained via model (13), after first verifying non-significance of a *Y***C* interaction term. These estimates are as follows: $\widehat{\pi}^*_{11}=0.272$, $\widehat{\pi}^*_{10}=0.178$, $\widehat{\pi}^*_{01}=0.233$, $\widehat{\pi}^*_{00}=0.150$.

Table IV provides definitions and estimates for the predictive value weights in this example, based on two distinct sets of assumed $\mathrm{SE}_{y\mathbf{c}}$ and $\mathrm{SP}_{y\mathbf{c}}$ values. The first case assumes non-differentiality with respect to both *Y* and *C*, such that $\mathrm{SE}_{y\mathbf{c}}$ =0.7, $\mathrm{SP}_{y\mathbf{c}}$ =0.9 (*y* =0, 1;*c* =0,1). The second illustrates a differential case in which SE and SP are allowed to vary with *Y* and *C*, as follows: SE_{11} =0.75, SE_{10} =0.75, SE_{01} =0.8, SE_{00} =0.8, SP_{11} =0.85, SP_{10} =0.85, SP_{01} =0.9, SP_{00} =0.9. The natïve estimates and standard errors for the coefficients corresponding to *X* and *C* (β and γ) are 0.206 (0.152) and 0.497 (0.126), respectively. This yields natïve adjusted OR estimates of 1.22 (*X*) and 1.64 (*C*). In contrast, in the non-differential case the predictive value weighting approach yields $\widehat{\beta}=0.406\ \ (0.318)$ and $\widehat{\gamma}=0.456\ \ (0.134)$, corresponding to adjusted OR estimates of 1.50 and 1.58 (standard errors obtained via the jackknife). In the differential case, we obtain $\widehat{\beta}=-0.156\ \ (0.415)$ and $\widehat{\gamma}=0.538\ \ (0.140)$, corresponding to adjusted OR estimates of 0.86 and 1.71. Note here that non-differential misclassification implies attenuation of the naïve β estimate, whereas the assumed differential misclassification pattern yields a naïve estimate on the opposite side of the null relative to the adjusted estimate. The estimates of γ are also impacted by misclassification of *X*, though more modestly than those of β.

3.3. Adjusting for multiple and continuous covariates

In this section, we use an example data set from the logistic regression text by Hosmer and Lemeshow [28]. The original data on 380 subjects with prostate cancer were obtained at The Ohio State University Comprehensive Cancer Center, and were modified to protect confidentiality. We utilize data from 378 patients (two men with Gleason score values of 0 were removed). The outcome (*Y*) is a binary indicator for whether or not cancer had penetrated the prostatic capsule, and the predictor variables that we consider are the patient's Gleason score and prostate specific antigen (PSA) value (C_1 and C_2; both continuous), and a binary indicator (X, but observed as Z) for whether nodules were detected via a digital rectal exam. For illustrative purposes, we assume that nodule detection was subject to misclassification.

As described in Section 2.4, estimated $\pi^*_{y\mathbf{c}}$ values were obtained via model (13), with Z as the outcome and Y, C_1, and C_2 as covariates. The $\widehat{\pi}^*_{y\mathbf{c}}$'s ranged from 0.52 to 0.93, with 22 values exceeding 0.9. Table V provides the 'naïve' logistic regression results, together with the results of applying the ML and predictive value weighting methods under the assumption that nodule detection was non-differentially misclassified with SE=0.9 and SP=0.8. We performed both analyses with and without adjusting the assumed value of $\text{SE}_{y\mathbf{c}}$ upward to equal $\widehat{\pi}^*_{y\mathbf{c}}$ for the 22 subjects with $\widehat{\pi}^*_{y\mathbf{c}}>0.9$ [see equations (15)]. The results in Table V are based on making this adjustment in conjunction with both the ML and predictive value weighting methods, but in both cases they differed only slightly from the unadjusted results. Note that accounting for the misclassification in Z as a surrogate for X markedly moves the naïve OR estimate for nodules away from the null (e.g. $e^{2.348}$ =10.46 via weighting, vs the naïve estimate $e^{1.172}$ =3.23). Also note that the ML estimates are accompanied by smaller standard errors than those based on predictive value weighting in this example. However, simulation studies (see next section) indicate that this result may not reflect general trends.

Table VI gives predictive value weighting-based estimated log ORs corresponding to true nodule status (X), along with jackknife standard errors, for several assumed (SE, SP) combinations in the prostate cancer example. These results suggest that in this case ln(OR) is much more sensitive to changes in SE given a fixed SP than to changes in SP for a fixed SE, an opposite conclusion to that found in our no-covariate example (Section 3.1; Figure 1). Though not shown in the table, estimated ln(OR)'s corresponding to C_1 and C_2 varied only slightly with the assumed (SE, SP) values, and were similar to the 'naïve' estimates for those variables.

3.4. Placing an assumed distribution on (SE, SP)

In previous examples, we calculated standard errors assuming fixed and known investigator-supplied (SE, SP) values. Some authors [10--14] suggest building in additional variability due to uncertainty about these misclassification parameters by specifying underlying joint densities for them and applying imputation-like or Bayesian methods. Such accommodation is readily made using the approach advocated here, although the necessity may be debatable (see Discussion).

To illustrate, we revisit the SIDS example (Table I; Section 3.1). Suppose that we presume non-differential misclassification of exposure, and wish to summarize uncertainty about SE and SP by assuming that they each derive independently from a trapezoidal distribution (see, e.g. [11]) with a minimum of 0.85, maximum of 1, and lower and upper modes of 0.9 and 0.95.

In this case, the goal of sensitivity analysis is to produce one point estimate of the odds ratio, together with an interval estimate that simultaneously takes account of the variability in the observed data (Table I) and the postulated systematic variability of SE and SP. We accomplished this as follows. First, we independently selected 2000 (SE, SP) pairs, with each taken randomly from the assumed trapezoidal distribution. For each pair, we computed the estimated ln(OR) via the predictive value weighting method, together with its jackknife standard error. We then generated 500 random draws from a normal distribution with mean and standard deviation matching that estimated ln(OR) and its associated standard error. Repeating this process for each of the 2000 (SE, SP) pairs and pooling the results produced a histogram of 2000×500 ln(OR) values, which is depicted in Figure 2. The 2.5th, 50th, and

97.5th percentiles of this distribution are −0.053, 0.418, and 1.155, respectively. Exponentiating these produces a median OR estimate of 1.52, with approximate 95 per cent confidence limits of (0.95, 3.17). These may be contrasted with the 'naïve' estimates of 1.31 (0.98, 1.71). This approach to interval estimation accounting for both random and systematic variation is akin to that suggested by Fox *et al*. [11], except that we produced each of our 2000 ln(OR) estimates via a single logistic regression with predictive value weighting, rather than by simulation.

4. Simulation studies

We conducted numerous simulation studies to validate point estimates produced by the predictive value weighting method, and to assess the appropriateness of the corresponding jackknife standard errors. Here we first report the results of one such study, assuming one binary covariate as in the example described in Section 3.2. A total of 1000 sets of data were generated under model (9), with two binary predictors (X and C) and a total sample size of 1144 (to mimic the SIDS example). The misclassified variable Z was generated assuming the following differential conditions: $SE_{11}=SP_{11}=0.7$, $SE_{10}=SP_{10}=0.75$, $SE_{01}=SP_{01}=0.8$, $SE_{00}=SP_{00}=0.85$. The true coefficients under model (9) were selected as follows: $\alpha=-0.8$, $\beta=1.7$, and $\gamma=1$. The prevalences of X and C and their association were determined by the following simulation conditions: $\Pr(X=1)=0.5$, $\Pr(C=1|X=1)=0.66$, and $\Pr(C=1|X=0)=0.41$.

Table VII summarizes the results of this first simulation study. Under these conditions, fitting the 'naïve' logistic regression that inserts Z in place of the unknown true binary exposure X leads to marked attenuation in the estimates of both β and γ, with severely subnominal confidence interval (CI) coverage (0 per cent in the case of β). In contrast, the estimates of β and γ based on the ML approach [equation (10)] and the predictive value weighting [equation (12)] approach are nearly identical and display minimal small-sample bias. Both methods produce average standard errors close to the empirical standard deviations of the point estimates, and yield excellent CI coverage. In particular, we find the jackknife standard error estimate to be quite reliable and accurate in conjunction with predictive value weighting.

Our second simulation study is designed to incorporate continuous covariates by mimicking the conditions of the prostate cancer example. Specifically, 500 independent data sets of size $n = 378$ were generated via models (9) and (13), with true parameters matching those estimated for the prostate cancer data via the ML method with the adjustment in (15) (see Table V). For each data set, we applied the ML and predictive value weighting methods, with and without adjustment. The results based on simulations incorporating adjustment are summarized in Table VIII, because (particularly in the case of ML) we observed more numerical problems without it. With regard to numerical stability, only 2 of 500 simulations led to 'blow ups' of the estimate of β and/or its standard error in using predictive value weighting, whereas 24 of 500 simulations reflected such problems based on ML. The results in Table VIII are based on the remaining simulated data sets in each case (see table footnote).

Other key features to note in Table VIII are the marked attenuation of the naïve estimate for β, along with upward bias in the naïve estimate for γ_2. The ML and weighting methods perform similarly, although there are more pronounced differences than were seen in the binary covariate case (Table VII). In particular, we observe slight positive bias in the ML estimator for β, and negative bias of a similar magnitude based on predictive value weighting. Confidence interval coverage appears slightly better for ML, possibly in part due to larger mean estimated standard errors. The methods perform almost identically when estimating γ_1 and γ_2.

5. Discussion

We have attempted to provide a relatively complete treatment of sensitivity analysis for binary covariate and outcome misclassification in the logistic regression setting. This topic has been emphasized in recent epidemiological literature, and our approach to covariate misclassification appears to have close connections with the macro implementations of Fox *et al*. [11]. However, the predictive value weighting approach provides added flexibility and crystallizes a technique that was previously only possible via simulation. We have compared that approach with direct ML in order to make informed recommendations for practical use. We hope the details and examples provided here, together with evaluating the statistical performance of the methods, will promote further interest in these concepts among statisticians and epidemiologists.

Our study permits numerous insights into the process of sensitivity analysis. First, we distinguish the cases of outcome and covariate misclassification, finding the former to be readily addressed via direct likelihood parameterizations [equations (8) and (16)]. Note in particular that in that case, the likelihood relies only on the assumed (SE, SP) values and the underlying logistic model [e.g. (3) or (9)] of interest. For the simple 2×2 case with misclassified exposure, we find straightforward ‘matrix-method’-based [3] identities [equation (1)] to be a fully adequate basis for effective sensitivity analysis. Such identities also clarify restrictions [equations (2) and (14)] that, while not explicitly recognized in some prior references, may need to be observed in order to prevent numerical breakdowns when seeking adjusted parameter estimates.

For the case of exposure misclassification accounting for covariates, our primary contribution is the development of the predictive value weighting approach [equations (6) and (12)] and its comparison against ML [equation (4) and expression (10)]. Our simulation results suggest that these two methods perform nearly identically in the no-covariate case, and in the case of all categorical covariates. In the case of continuous covariates, however, we find that predictive value weighting can be more stable than ML via numerical optimization (see footnote to Table VIII). Another useful way to compare and contrast the two methods is in terms of the additional modeling that each requires [see comment following equation (13)].

Our main focus has been the process of obtaining estimated log odds ratios in logistic regression that are adjusted for misclassification based on assumed SE and SP parameters, together with valid standard errors that account for variation in the data actually observed. We believe this usually constitutes the primary aim of such sensitivity analyses. In Section 3.4, however, we illustrate the use of assumed distributions (essentially priors) for SE and SP, in order to find a single adjusted point estimate and confidence interval accounting for both random and systematic variation. This is the approach advocated by some authors [11, 12], from whom we borrow the idea of the trapezoidal distribution for illustration. In the case of differential misclassification, these authors discuss the use of assumed joint distributions for SE and SP that account for their possible correlation. These may be readily incorporated along with the predictive value weighting approach advocated here, although we recommend use of such assumed distributions only when it best matches overall objectives. For example, a sensitivity analysis summarizing results based on a collection of assumed (SE, SP) values, such as that illustrated in Figure 1, will sometimes be more informative. In that case, the valuable insight that the odds ratio is primarily sensitive to variations in SP rather than SE would have been masked by assuming a joint distribution for those parameters to produce a single point estimate.

The favorable performance of predictive value weighting relative to ML is encouraging, as the former also appears well-equipped to facilitate natural generalizations (Section 2.6). Perhaps the main advantage of predictive value weighting, however, is the intuitive appeal of the expanded data set with appropriate weights, which is analyzed using standard logistic regression software. This makes sensitivity analysis more computationally accessible, obviating the need for either simulation [10, 11] or numerical likelihood maximization to obtain point estimates. Variations of such case weighting approaches have proven useful for the analysis of incomplete data [29, 30], as well as in other contexts such as power and sample size approximation [31].

The ability to allow SE and SP parameters to vary according to subject characteristics (Section 2.4) is critical, given the realism of this scenario in practice [32, 33]. However, postulating appropriate values for $\mathrm{SE}_{y\mathbf{c}}$ and $\mathrm{SP}_{y\mathbf{c}}$ poses obvious challenges when multiple covariates are considered. Future work may identify a role for predictive value weighting in regression-based predictor variable misclassification corrections when internal validation data are available to identify these parameters.

Acknowledgments

Robert H. Lyles was partially supported by an R01 from the National Institute of Environmental Health Sciences (ES012458), and by a PHS grant (UL 1 RR025008) from the Clinical and Translational Science Award Program, National Institutes of Health, Center for Research Resources. We thank the editors and reviewers for insightful and constructive comments.

Contract/grant sponsor: National Institute of Environmental Health Sciences; contract/grant number: ES012458

Contract/grant sponsor: National Institutes of Health; contract/grant number: UL 1 RR025008

References

1. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models. 2nd edn. Chapman & Hall; New York: 2006.
2. Bross IDJ. Misclassification in 2×2 tables. Biometrics. 1954; 10:478–486.
3. Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics. 1977; 33:414–418. [PubMed: 884199]
4. Greeenland S, Kleinbaum D. Correcting for misclassification in two-way tables and matched-pair studies. International Journal of Epidemiology. 1983; 12:93–97. [PubMed: 6840961]
5. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. Statistics in Medicine. 1987; 7:745–757. [PubMed: 3043623]
6. Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. Journal of Clinical Epidemiology. 1990; 43:941–947. [PubMed: 2213082]
7. Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics. 1999; 55:338–344. [PubMed: 11318185]
8. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. Statistics in Medicine. 1986; 5:21–27. [PubMed: 3961312]
9. Liu X, Liang K-Y. Adjustment for nondifferential misclassification error in the generalized linear model. Statistics in Medicine. 1991; 10:1197–1211. [PubMed: 1925152]
10. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology. 2003; 14:451–458. [PubMed: 12843771]
11. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. International Journal of Epidemiology. 2005; 34:1370–1376. [PubMed: 16172102]

12. Chu H, Wang Z, Cole SR, Greenland S. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. Annals of Epidemiology. 2006; 16:834–841. [PubMed: 16843678]

13. Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. Stata Journal. 2008; 8:29–48.

14. Gustafson P, Le ND, Saskin R. Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics. 2001; 57:598–609. [PubMed: 11414590]

15. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology. 1997; 146:195–203. [PubMed: 9230782]

16. Drews CD, Kraus JF, Greenland S. Recall bias in a case-control study of sudden infant death syndrome. International Journal of Epidemiology. 1990; 19:405–411. [PubMed: 2376455]

17. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure–disease relationships and methods of correction. Annual Reviews of Public Health. 1993; 14:69–93.

18. Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics. 2002; 58:1034–1037. [PubMed: 12495160]

19. Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. Journal of Statistical Planning and Inference. 2008; 138:528–538.

20. SAS Institute, Inc.. SAS/STAT 9.1 User's Guide. Vol. vol. 4. SAS Institute; Cary, NC: 2004.

21. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.

22. Hinkley, D. Jackknife methods. In: Kotz, S.; Johnson, NL., editors. Encyclopedia of Statistical Sciences. Vol. vol. 4. Wiley; New York: 1983. p. 280-287.

23. Neuhaus JM. Bias and efficiency loss due to misclassified responses in logistic regression. Biometrika. 1999; 86:843–855.

24. Neuhaus JM. Analysis of clustered and longitudinal binary data subject to response misclassification. Biometrics. 2002; 58:675–683. [PubMed: 12230004]

25. SAS Institute, Inc.. SAS IML 9.1 User's Guide. SAS Institute; Cary, NC: 2004.

26. Venables, WN.; Ripley, BD. Modern Applied Statistics with S-Plus. Springer; New York: 1994.

27. Everitt, BS.; Hothorn, T. A Handbook of Statistical Analyses Using R. Chapman & Hall; New York: 2006.

28. Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. 2nd edn. Wiley; New York: 2000.

29. Fleiss, JL.; Levin, B.; Paik, MC. Statistical Methods for Rates and Proportions. 3rd edn. Wiley; New York: 2003.

30. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika. 1996; 83:916–922.

31. Lyles RH, Lin H-M, Williamson JM. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statistics in Medicine. 2007; 26:1632–1648. [PubMed: 16817148]

32. Hlatky M, Pryor D, Harrell F, Califf R, Mark D, Rosati R. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. The American Journal of Medicine. 1984; 77:64–71. [PubMed: 6741986]

33. Karakaya J, Aksoy D, Harmanci A, Karaagaoglu E, Gurlek A. Predictive ability of fasting plasma glucose for a diabetic 2-h postload glucose value in oral glucose tolerance test: spectrum effect. Journal of Diabetes and Its Complications. 2007; 21:300–305. [PubMed: 17825754]

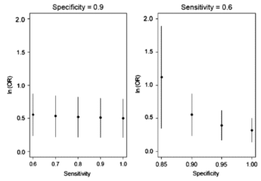

Figure 1.

Point ln(OR) estimates with error bars (±1 jackknife standard error) for several (SE, SP) combinations assuming non-differential misclassification of exposure in conjunction with the SIDS study data in Table I.

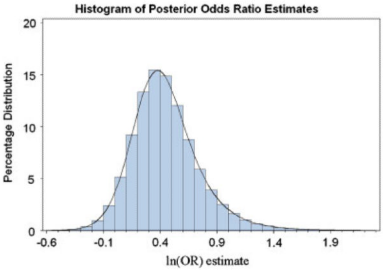

Figure 2.

Histogram of 2000×500 ln(OR) estimates accounting for both random error in the observed SIDS study data (Table I) and systematic error in the assumed distributions of SE and SP. Here, SE and SP are assumed to be derived independently from the trapezoidal distribution described in Section 3.4. The vertical axis represents the percentage, and a kernel smooth is overlaid.

Table I

Data from case-control study of sudden infant death syndrome (SIDS) and reported maternal antibiotic use.[*]

	Self-reported use of antibiotics (*Z*)	
Case-control status (*Y*)	**1**	**0**
1	122	442
0	101	479

[*] Reference: Drews *et al*. [16].

Table II

Expanded data set for example based on Table I.

y	z	x	Weight (w_{xzy})	# obs. per Table I	w_{xzy}'s for Table I data:*	
					Non-differential	Differential
1	1	1	w_{111}=Pr(X =1|Z =1, Y =1)=PPV$_1$	122	0.645	0.113
1	1	0	w_{011}=1−PPV$_1$	122	0.355	0.887
1	0	1	w_{101}=Pr(X =1|Z =0, Y =1)=1−NPV$_1$	442	0.119	0.021
1	0	0	w_{001}=NPV$_1$	442	0.881	0.979
0	1	1	w_{110}=Pr(X =1|Z =1, Y =0)=PPV$_0$	101	0.511	0.497
0	1	0	w_{010}=1−PPV$_0$	101	0.489	0.503
0	0	1	w_{100}=Pr(X =1|Z =0, Y =0)=1−NPV$_0$	479	0.072	0.045
0	0	0	w_{000}=NPV$_0$	479	0.928	0.955

*Non-differential case: (SE,SP)=(0.6,0.9); differential: (SE_1,SP_1,SE_0,SP_0)=(0.6,0.8,0.7,0.9).

Table III

Data from case-control study of sudden infant death syndrome (SIDS) and reported maternal antibiotic use.*

	$C = 1$†		$C = 0$†	
	Self-reported use of antibiotics (*Z*)		**Self-reported use of antibiotics (*Z*)**	
Case-control status (*Y*)	**1**	**0**	**1**	**0**
1	60	170	62	272
0	42	127	59	352

*Reference: Drews *et al*. [16].

†Artificial binary stratification variable (*C*) used for illustration

Table IV

Expanded data set for example based on Table III.

y	z	c	x	Weight	# obs. per Table III	Weights for Table III data:*	
						Non-differential	Differential
1	1	1	1	Pr(X =1|Z =1, Y =1, C =1)=PPV_{11}	60	0.738	0.561
1	1	1	0	1−PPV_{11}	60	0.262	0.439
1	1	0	1	Pr(X =1|Z =1, Y =1, C =0)=PPV_{10}	62	0.511	0.196
1	1	0	0	1−PPV_{10}	62	0.489	0.804
1	0	1	1	Pr(X =1|Z =0, Y =1, C =1)=1−NPV_{11}	170	0.118	0.070
1	0	1	0	NPV_{11}	170	0.882	0.930
1	0	0	1	Pr(X =1|Z =0, Y =1, C =0)=1−NPV_{10}	272	0.047	0.014
1	0	0	0	NPV_{10}	272	0.953	0.986
0	1	1	1	Pr(X =1|Z =1, Y =0, C =1)=PPV_{01}	42	0.667	0.653
0	1	1	0	1−PPV_{01}	42	0.333	0.347
0	1	0	1	Pr(X =1|Z =1, Y =0, C =0)=PPV_{00}	59	0.388	0.380
0	1	0	0	1−PPV_{00}	59	0.612	0.620
0	0	1	1	Pr(X =1|Z =0, Y =0, C =1)=1−NPV_{01}	127	0.087	0.050
0	0	1	0	NPV_{01}	127	0.913	0.950
0	0	0	1	Pr(X =1|Z =0, Y =0, C =0)=1−NPV_{00}	352	0.029	0.017
0	0	0	0	NPV_{00}	352	0.971	0.983

[*] Non-differential case: (SE_{11},SE_{10},SE_{01},SE_{00},SP_{11},SP_{10},SP_{01},SP_{00})=(0.7,0.7,0.7,0.7,0.9,0.9,0.9,0.9); differential: (SE_{11},SE_{10},SE_{01},SE_{00},SP_{11},SP_{10},SP_{01},SP_{00})=(0.75,0.75,0.8,0.8,0.85,0.85,0.9,0.9).

Table V

‘Naïve’ and misclassification-adjusted results for prostate cancer data example.[*]

	‘Naïve’ model **Estimate (std. error)**	**Maximum likelihood** [†] **Estimate (std. error)**	**Predictive value weighting** [‡] **Estimate (std. error)**
Intercept	−8.197 (1.055)	−8.708 (1.295)	−8.773 (1.454)
Presence of nodules (*X*)	1.172 (0.322)	2.160 (0.842)	2.348 (1.039)
PSA (C_1)	0.029 (0.009)	0.029 (0.010)	0.030 (0.012)
Gleason score (C_2)	1.000 (0.161)	0.936 (0.172)	0.918 (0.178)

[*] Data example from Hosmer and Lemeshow [28]; Outcome (*Y*)=prostatic capsule penetration; assuming nodule detection misclassified with SE=0.9, SP=0.8.

[†] See expression (10); adjustment in (15) applied to 22 of the 378 subjects.

[‡] See equation (12); std. errors via jackknife; adjustment in (15) applied to 22 of the 378 subjects.

Table VI

‘Naïve’ and non-differential misclassification-adjusted results for prostate cancer data example, assuming a range of (SE, SP) values.[*],[†]

	Assumed SE	
Assumed SP	**0.9**	**0.95**
0.7	2.449 (1.072)	1.666 (0.488)
0.8	2.348 (1.039)	1.581 (0.467)
0.9	2.274 (1.014)	1.520 (0.451)

[*] Data example from Hosmer and Lemeshow [28]; Outcome (*Y*)=prostatic capsule penetration.

[†] Table entries correspond to estimated adjusted coefficient for presence of nodules (*X*), via predictive value weighting method [see equation (12)]; std. errors via jackknife; adjustment in (15) applied to 22 of the 378 subjects when SE=0.9.

Table VII

Simulation results for logistic regression with *X* misclassified and one other binary covariate (*C*).*

Parameter	'Naïve' method	Maximum likelihood method†	Weighting method‡
β	0.93 (0.14)	1.72 (0.28)	1.72 (0.28)
	[0.13]	[0.28]	[0.28]
	{0.0 per cent}	{94.3 per cent}	{94.6 per cent}
γ	1.19 (0.13)	0.99 (0.17)	0.97 (0.17)
	[0.13]	[0.17]	[0.18]
	{70.5 per cent}	{96.5 per cent}	{97.5 per cent}

*Numbers in each cell reflect mean (standard deviation) based on 1000 simulated data sets, with true values β=1.7 and γ=1.0. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95 per cent confidence interval coverage rates.

†See expression (10).

‡See equation (12); std. errors via jackknife; assumed SE and SP values given in text of Section 4.

Table VIII

Simulation results for logistic regression with X misclassified and two continuous covariates (C_1 and C_2).[*]

Parameter	'Naïve' method	Maximum likelihood method[†]	Weighting method[‡]
β	1.16 (0.34)	2.29 (0.88)	2.06 (0.87)
	[0.33]	[1.22]	[0.99]
	{16.0 per cent}	{94.5 per cent}	{93.6 per cent}
γ_1	0.03 (0.009)	0.03 (0.010)	0.03 (0.010)
	[0.009]	[0.010]	[0.010]
	{95.0 per cent}	{95.8 per cent}	{95.0 per cent}
γ_2	1.03 (0.12)	0.96 (0.13)	0.92 (0.13)
	[0.13]	[0.14]	[0.14]
	{92.0 per cent}	{97.1 per cent}	{95.4 per cent}

[*] Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets, with true values β=2.16, γ_1=0.03, and γ_2=0.94. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95 per cent confidence interval coverage rates.

[†] See expression (10); 24 of 500 runs discarded due to numerical instability.

[‡] See equation (12); std. errors via jackknife; assumed SE and SP values given in text of Section 4; 2 of 500 runs discarded due to numerical instability.