STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2221–2234
Published online 11 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2094

Does it always help to adjust for misclassification of a binary outcome in logistic regression?

Xianqun Luan,[1] Wei Pan,[2,‡] Susan G. Gerberich[3] and Bradley P. Carlin[2,*,†]

[1]*Division of Biostatistics and Epidemiology, The Children's Hospital of Philadelphia, 3535 Market St., 14th Floor, Philadelphia, PA 19104, U.S.A.*
[2]*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455-0378, U.S.A.*
[3]*Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN 55455-0378, U.S.A.*

SUMMARY

It is well known that in logistic regression, where the outcome is measured with error, a biased estimate of the association between the outcome and a risk factor may result if no proper adjustment is made. Hence, it seems tempting to always adjust for possible misclassification of the outcome. Here we show that it is not always beneficial to do so because, though the adjustment reduces the bias, it also inflates the variance, leading to a possibly larger mean squared error of the estimate. In the context of a data set on agricultural injuries, numerical evidence is provided through simulation studies. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: bias; logistic models; mean squared error; measurement error; sensitivity and specificity; simulation

INTRODUCTION

Misclassification of a binary outcome refers to the measurement error in the outcome [1]. In the regional rural injury study II (RRIS-II), the outcome variable, agricultural injury status (injured or not injured), was collected based on telephone interview and might be incorrectly reported. According to an external validation study, the Olmsted agricultural trauma study

*Correspondence to: Bradley P. Carlin, Division of Biostatistics, University of Minnesota, School of Public Health, Box 303 Mayo Memorial Building, 420 Delaware St., Minneapolis, MN 55455-0378, U.S.A.
†E-mail: brad@biostat.umn.edu
‡E-mail: weip@biostat.umn.edu

Contract/grant sponsor: NIH; contract/grant number: R01-HL65462
Contract/grant sponsor: NIAID; contract/grant number: R01-AI41966
Contract/grant sponsor: NSF/EPA; contract/grant number: SES 99-78238

Received March 2002
Accepted September 2004

(OATS) [2, 3], the specificity and sensitivity of the outcome classification were 0.983 and 0.689, respectively; that is, 1.7 per cent of persons not having incurred agricultural injuries and 31.1 per cent of persons having incurred such injuries were not classified according to their actual statuses. Misclassification might be due to several reasons, including recall bias, or missing information (as when a respondent ignored some minor injuries, or did not know about all injuries that were incurred by each member of the household).

It is well known that such misclassification can result in biased estimates of the association between the outcome and relevant covariates. In this paper we consider logistic regression when the binary outcome may be misclassified. If the misclassification is *non-differential* (i.e. the misclassification does not depend on other variables), estimates are biased toward the null value [4]. If we know or can estimate possibly individual-specific sensitivity and specificity, an expectation–maximization (EM) algorithm can be used to adjust for misclassification; see Reference [4] for details. Since the adjustment method is easy to implement, it may appear appropriate to simply always do so if misclassification of the outcome is suspected. However, in this paper, we show through numerical studies that the issue is much more complicated: whether to adjust may depend on several factors, including for instance the sample size and the primary question of interest.

It is known that there is a bias-variance trade-off in many measurement error contexts (see e.g. Reference [5, p. 32]). Although an adjustment method can reduce the bias, it will often increase the variance. The mean squared error (MSE), the sum of the squared bias and the variance, is probably the most common criterion used to compare point estimators. Hence, if the primary question is to estimate the strength of association between the outcome and a covariate (say, an exposure), the bias-corrected estimator may actually perform less well (i.e. have larger MSE) than the uncorrected one. However, because the variance usually decreases as the sample size increases, the bias-corrected estimator will always have the smaller MSE if the sample size is large enough. On the other hand, if the primary goal is to construct confidence intervals or hypothesis tests with appropriate nominal levels, it is often necessary to correct the bias. These points will be illustrated through simulation studies whose generated data mimic that of RRIS-II [6, 3] and OATS. We will also show that the performance of adjustment methods is unsatisfactory if they do not take into account any differential misclassification that is present.

AGRICULTURAL INJURY STUDIES

Injury study: RRIS-II

The RRIS-II consisted of two parts: a cross-sectional study and a case-control study. Farm/ranch households were randomly selected from five states: Minnesota (MN), Nebraska (NE), South Dakota (SD), North Dakota (ND) and Wisconsin (WI), using the United States Department of Agriculture (U.S.DA) National Agricultural Statistics Services (NASS) Master Sampling Frame of Farm Operations. Demographic and injury information for all household members were collected in the cross-sectional portion of the study by computer assisted telephone interviews (CATI). Information for 3781 households was obtained. For the purpose of the current study, one person was randomly selected from each household; the 94 households with any missing data were removed. As a result, a total of 3687 households were included

Table I. Univariate analysis of agricultural injuries

	Non-injured		Injured			
Variables	*N*	per cent	*N*	per cent	Total	*P*
Residence states						0.4265
MN	668	97.23	19	2.77	687	
NE	735	96.71	25	3.29	760	
ND	766	95.75	34	4.25	800	
SD	767	95.64	35	4.36	802	
WI	616	96.55	22	3.45	638	
Total	3552	96.34	135	3.66	3687	
Age group						<0.0001
0–9	624	98.42	10	1.58	634	
10–19	1123	98.6	16	1.4	1139	
20–39	706	94.64	40	5.36	746	
40–59	1064	94.08	67	5.92	1131	
60 +	35	94.59	2	5.41	37	
Total	3552	96.34	135	3.66	3687	
Gender						<0.0001
Female	1691	97.8	38	2.2	1729	
Male	1861	95.05	97	4.95	1958	
Total	3552	96.34	135	3.66	3687	
Farm work time (h/6 months)						<0.0001
None	872	99.09	8	0.91	880	
1–249	991	98.41	16	1.59	1007	
250–499	470	97.51	12	2.49	482	
500–999	469	96.3	18	3.7	487	
1000–1499	321	88.92	40	11.08	361	
1500–1999	287	91.11	28	8.89	315	
2000 +	142	91.61	13	8.39	155	
Total	3552	96.34	135	3.66	3687	

in the final study sample. The target period for identifying the injury outcome in this analysis was January 1, 1999 to June 30, 1999.

The rates of agricultural injuries with respect to demographic factors and farm work time are summarized in Table I. The overall injury rate was 3.7 per cent. This rate was not significantly different among the five states ($P = 0.4265$). However, age group was significantly associated with injury ($P < 0.0001$), with injury rate tending to increase with age. For example, the injury rate is about 1.5 per cent for persons 19 years or younger, compared with more than 5 per cent for persons 20 years or older. Injury rates also differed significantly by gender ($P < 0.0001$): females accounted for 46.9 per cent of the study sample, but only 28.1 per cent of persons having incurred farm-related injuries. Finally, the injury rate was significantly associated with farm work time (representing the degree of exposure to agricultural activities; $P < 0.0001$). The proportion injured varied between 0.91 per cent for persons who reported no farm work

during the six-month period, and 11.1 per cent for those whose reported farm work time of 1000–1499 h (Table I).

Generalized additive models [7] were employed to explore functional forms for age and farm work time. A linear term for age and a log transformation of farm work time ('lghrs') were selected. Note that 1 h was added to each person's farm work time before making the log transformation to avoid having a logarithm of zero. In our variable selection process, state of residence did not emerge as significant, and thus was dropped from the model. While gender was not statistically significant, it was retained due to its substantive importance in our model; both age and lghrs were highly significant. The fitted model was

$$\text{logit}(\pi) = -5.584 + 0.161\ \text{Gender} + 0.022\ \text{Age} + 0.288\ \text{lghrs} \tag{1}$$

where π is the probability of incurring a farm-related injury, Gender $= 1$ for males and -1 for females, and Age was measured in years. Standard statistical checks of this model revealed no evidence of lack of fit.

Validation study: OATS

The validation study [2] was actually a sub-study of the OATS [3], which included all farms in Olmsted County, Minnesota identified in the U.S.DA NASS master sampling frame. Data were collected using telephone interviews. A medical record validation process using the Mayo Clinic's medical record system (Rochester Epidemiology Project), which included health care records for virtually all residents of Olmsted County, Minnesota, was employed to validate the telephone interview-reported injuries and identify any injuries among two study samples with separate one year time (from November 1, 1986 to October 31, 1987, and from June 1, 1987 to May 31, 1988) but were not reported during the interview.

There were several variables that were commonly present in both the telephone interviews and the medical records. These variables were examined in the validation analyses conducted for the current study. They included whether the injury was agriculturally related or not, age and gender of the subject, the anatomical site, type, and source of the injury, the length of time between the injury and the interview, and the key respondent involved. Analyses were conducted to determine the agreement between the telephone interview injury data and the data from the medical records. In our context, sensitivity is the extent to which persons who truly have incurred agriculturally related injuries are so classified, while specificity is the extent to which persons who have not incurred farm-related injuries are so classified. For simplicity, only subjects with farm-related injury events and subjects without any injury events were included, while subjects with only non-farm-related injury events and those with events that could not be classified as either farm or non-farm were excluded from the analyses.

The sensitivity and specificity of the classification of farm-related injury status based on telephone interviews in RRIS-II were estimated from the OATS [2, 3]. Table II shows the result if we treat both the sensitivity and specificity as constant. The estimated specificity is 0.983 with 95 per cent CI (0.975, 0.991). The estimated sensitivity is 0.689 with 95 per cent CI (0.573, 0.805).

To investigate whether the sensitivity or specificity was associated with other variables, including age, gender, lghrs, and the type of individual who responded to the telephone interview (Resptype $= 0$ if the respondent was the female head of household, Resptype $= 1$ if the respondent was the male head of household, and Resptype $= 2$ otherwise), logistic

Table II. Classification of agricultural injury status.

		Validated classification		
		Non-injured	Injured	Total
Reported classification	Non-injured	964	19	983
	Injured	17	42	59
Total		981	61	1042

regression was employed. It was found that the sensitivity was not associated with any of the above variables; however, the specificity was significantly associated with Resptype and lghrs. Misclassification was more likely to occur with a male respondent than a female respondent (OR = 5.03, 95 per cent CI (1.84, 13.72)). A person who did not incur farm-related injuries was more likely to be misclassified when he/she spent more time on farm-related work (OR = 1.67, 95 per cent CI (1.34, 2.09)). The fitted model was

$$\text{logit}(\text{specificity}) = 7.920 - 1.615(\text{Resptype} = 1) - 1.380(\text{Resptype} = 2) - 0.515\,\text{lghrs} \quad (2)$$

Because the term for (Resptype = 2) was not statistically significant (with p-value = 0.2096), it would be dropped out in the following simulation study.

METHODS

The Magder–Hughes Procedure

Suppose that for subject i, T_i is the true binary outcome, Y_i is the observed (binary) outcome that is possibly misclassified, and $X_i = (X_{i1}, \ldots, X_{ik})$ is the set of covariates. As usual, T_i and Y_i are coded as 1 or 0, representing an event (e.g. a farm injury in RRIS-II) or no event for subject i. A logistic regression model is assumed to be

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} \quad (3)$$

where $\pi_i = \Pr(T_i = 1 | X_i)$.

To obtain unbiased estimates, Magder and Hughes proposed the following adjustment procedure. First, we have to assume that we know the sensitivity

$$\text{sens}_i = \Pr(Y_i = 1 | T_i = 1)$$

and specificity

$$\text{spec}_i = \Pr(Y_i = 0 | T_i = 0)$$

for each subject i. As shown earlier, the sensitivity and specificity can be estimated from a validation study such as OATS.

Let $\hat{Y}_i = \Pr(T_i = 1 | Y_i)$. By Bayes' theorem, we have that, for $Y_i = 1$,

$$\hat{Y}_i = \frac{\pi_i \text{sens}_i}{\pi_i \text{sens}_i + (1 - \pi_i)(1 - \text{spec}_i)}$$

and for $Y_i = 0$,

$$\hat{Y}_i = \frac{\pi_i(1 - \text{sens}_i)}{\pi_i(1 - \text{sens}_i) + (1 - \pi_i)\text{spec}_i}$$

To find maximum likelihood estimates of β_i's, an iterative procedure can be followed: (i) based on some initial or current values of β_i's, calculate $\hat{Y}_i$'s; (ii) perform a standard logistic regression with each subject classified as both with an event and without an event, with weights equal to $\hat{Y}_i$ and $1 - \hat{Y}_i$, respectively, leading to updated estimates of β_i's; (iii) repeat (i) and (ii) until convergence. More formally, this procedure is an EM algorithm. Variance estimates of β_i estimates can be also obtained.

Methods to be compared

Ideally, if T_i is observed, we will fit the above logistic regression model using T_i as the response variable. However, with possible misclassifications, we only observe Y_i but not T_i. A simple method would be to use Y_i as the response variable in the logistic regression model, which will introduce biased estimates for β_i's (as to be shown later). Alternatively, if there is no differential misclassification, we could treat sens_i and spec_i as two constants, or instead use subject-specific sens_i or spec_i. We wish to compare the performance of these various methods via simulation.

In what follows, we refer to the method that does not adjust for misclassification as the *naive* method. Here a standard logistic regression model is fit using the (possibly misclassified) outcomes Y_i. The second method, which we call *adj-1*, adjusts for misclassification but using a constant sensitivity $= 0.689$ and (incorrectly) a constant specificity $= 0.983$. The third method, called *adj-2*, adjusts for misclassification using a constant sensitivity $= 0.689$ and individual-specific specificity estimated from (2). Hence, only the adj-2 method is based on the correct modeling assumptions, whereas the other two are not (though all three methods use possibly misclassified outcome Y_i). For comparison, we also consider a method that is optimal (but impossible to implement in practice) which we call the *standard*, wherein the logistic model is fit using the true (but unobservable) outcome T_i.

SIMULATIONS

Simulation set-ups

To investigate the performance of various bias-adjustment methods in the context of RRIS-II/OATS, simulation studies that mimic the set-up of these two studies were conducted. We considered two general set-ups, a null case and a non-null case: the true outcome was simulated from either a null model

$$\text{logit}(\pi) = -3.27 \tag{4}$$

or from a non-null model, the fitted logistic model (1); subject-specific specificity for differential misclassification was realized through the fitted model (2). Specifically, to run our simulation study, $K = 200$ or 500 simulated data sets (for $n = 1000$ or 3687, respectively) were generated in the following way. At iteration k ($k = 1, \ldots, K$), n subjects were randomly

selected (without replacement) from the study sample. The probability π_i of the ith person having incurred farm-related injuries was computed from the fitted logistic model (3) or (1), depending on whether we were considering the null case or non-null case. The true outcome T_i (1 if individual i did incur farm-related injuries, and 0 if not) was then generated from a Bernoulli distribution, Bern(π_i). Depending on whether $T_i = 1$ or 0, the observed outcome Y_i (possibly misclassified) was then generated from a Bernoulli distribution with success probability equal to the sensitivity or 1 minus the specificity, where the sensitivity was fixed at 0.689 and the specificity was calculated according to model (2). Hence, the specificity was individual-specific; i.e. we had differential misclassification.

Each method was applied to data set k to obtain an estimate of the jth log odds ratio $\hat{\beta}_{j,k}$ for $k = 1, \ldots, K$ and $j = 1, \ldots, 4$ (corresponding to the 4 model terms, namely intercept, gender, age, and lghrs). We then calculated the empirical mean, variance and MSE of the estimator as

$$\text{Mean}_j = \sum_{k=1}^{K} \hat{\beta}_{j,k}/K, \quad \text{Var}_j = \sum_{k=1}^{K} (\hat{\beta}_{j,k} - \text{Mean}_j)^2/(K-1), \quad \text{MSE}_j = \text{Mean}_j^2 + \text{Var}_j$$

and the bias of the estimator as $\text{Bias}_j = \text{Mean}_j - \beta_j$. Based on each $\hat{\beta}_{j,k}$ and its standard error, we could also construct a 95 per cent CI and see whether or not it covered the true value of β_j. The empirical coverage percentage (CP) of the 95 per cent CI was then the proportion of the K CIs that covered the true β_j.

SIMULATION RESULTS

Non-null case

The results for sample sizes $n = 1000$ and 3687 are listed in Tables III and IV, respectively. We can see that the naive method without adjustment leads to biased estimates. However, we also see that the adj-1 method does not work well: in Table III, it yields even larger biases for the intercept, gender, and lghrs than does the naive method; its variances are also uniformly higher. It appears that the violation of the differential misclassification assumption in adjustment has severe consequences, leading to poor estimates. Table IV shows that this phenomenon also holds when the sample size is increased. As such, we do not consider adj-1 further.

The adjustment method that uses the correct differential misclassification assumption, adj-2, does eliminate most of the biases. However, though nearly unbiased, the estimates from adj-2 also have larger variability than those based on the naive method, leading to possibly larger MSEs. As the sample size is increased from 1000 to 3687 (Table IV), the performance of adj-2 catches up with that of the naive method. Coverage probabilities for adj-2 are also closer to the nominal 95 per cent level, though adj-2 MSEs for the intercept and gender terms are still slightly larger than those of the naive method.

Since predictive settings are the most common ones where interval estimates and hypothesis tests are not of primary interest (and thus the adjusted methods may be less appropriate), Tables III and IV also compare the predictive performance of the four methods. Specifically, note that (3) implies that the probability of an event is

$$\pi_i = \text{logit}^{-1}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

Table III. Simulation results with the non-null case and sample size $n = 1000$, based on 200 simulated data sets.

Covariate (j)	True$_j$	Mean$_j$	Bias$_j$	Variance$_j$	MSE$_j$	$\widehat{se}$(MSE$_j$)	CP$_j$ (per cent)
Method: *standard*							
Intercept	−5.5848	−5.7257	−0.1410	0.6456	0.6622	0.0768	96.5
Gender	0.1607	0.1901	0.0294	0.0549	0.0554	0.0050	94.5
Age	0.0217	0.0225	0.0007	0.0002	0.0002	0.0000	94.5
lghrs	0.2878	0.2937	0.0059	0.0207	0.0206	0.0024	91.5
π_{1639}	0.0503	0.0489	−0.0014	0.0003	0.0003	0.0000	—
π_{1050}	0.1199	0.1209	0.0011	0.0010	0.0010	0.0001	—
π_{2690}	0.1479	0.1579	0.0100	0.0048	0.0049	0.0006	—
Method: *naive*							
Intercept	−5.5848	−5.7274	−0.1427	0.6518	0.6689	0.0858	95.5
Gender	0.1607	0.1249	−0.0357	0.0485	0.0495	0.0045	92.0
Age	0.0217	0.0148	−0.0069	0.0002	0.0002	0.0000	87.0
lghrs	0.2878	0.3607	0.0730	0.0209	0.0261	0.0031	94.0
π_{1639}	0.0503	0.0584	0.0080	0.0004	0.0005	0.0001	—
π_{1050}	0.1199	0.1230	0.0031	0.0010	0.0010	0.0001	—
π_{2690}	0.1479	0.1390	−0.0088	0.0037	0.0038	0.0004	—
Method: *adj*-1							
Intercept	−5.5848	−8.4949	−2.9101	5.0429	13.4840	1.6512	99.4
Gender	0.1607	0.2860	0.1253	0.2405	0.2549	0.0345	97.2
Age	0.0217	0.0361	0.0144	0.0007	0.0009	0.0001	95.6
lghrs	0.2878	0.6027	0.3149	0.1167	0.2152	0.0277	98.9
π_{1639}	0.0503	0.0459	−0.0045	0.0008	0.0008	0.0001	—
π_{1050}	0.1199	0.1830	0.0631	0.0035	0.0075	0.0007	—
π_{2690}	0.1479	0.2570	0.1091	0.0205	0.0323	0.0041	—
Method: *adj*-2							
Intercept	−5.5848	−5.8175	−0.2328	1.6459	1.6915	0.3864	95.8
Gender	0.1607	0.2485	0.0878	0.1262	0.1333	0.0150	96.9
Age	0.0217	0.0305	0.0087	0.0004	0.0005	0.0001	96.4
lghrs	0.2878	0.2451	−0.0427	0.0413	0.0429	0.0074	89.6
π_{1639}	0.0503	0.0457	−0.0046	0.0006	0.0006	0.0001	—
π_{1050}	0.1199	0.1298	0.0099	0.0023	0.0023	0.0002	—
π_{2690}	0.1479	0.1989	0.0511	0.0125	0.0151	0.0019	—

a 1-1 function of the β's given particular values of the covariates. Thus to compare the relative predictive performance of the methods, we have added lines to Tables III, IV corresponding to the π_i for three representative subjects: the first (subject 1639) with covariate values such that the true $\pi_i \approx 0.05$ (a relatively low value), the second (subject 1050) having true $\pi_i \approx 0.12$ (moderate), and the third (subject 2690) having true $\pi_i \approx 0.15$ (high). The results are very similar to those previously reported; namely, superior performance by the naive method in Table III, but only marginally better performance in Table IV.

Table IV. Simulation results with the non-null case and sample size $n = 3687$, based on 500 simulated data sets.

Covariate (j)	True$_j$	Mean$_j$	Bias$_j$	Variance$_j$	MSE$_j$	$\widehat{se}$(MSE$_j$)	CP$_j$ (per cent)
Method: *standard*							
Intercept	−5.5848	−5.6026	−0.0178	0.0784	0.0785	0.0049	94.6
Gender	0.1607	0.1701	0.0094	0.0056	0.0057	0.0004	94.0
Age	0.0217	0.0219	0.0002	0.0000	0.0000	0.0000	95.8
lghrs	0.2878	0.2880	0.0002	0.0024	0.0024	0.0002	92.8
π_{1639}	0.0503	0.0498	−0.0005	0.0000	0.0000	0.0000	—
π_{1050}	0.1199	0.1208	0.0009	0.0001	0.0001	0.0000	—
π_{2690}	0.1479	0.1502	0.0023	0.0005	0.0005	0.0000	—
Method: *naive*							
Intercept	−5.5848	−5.6459	−0.0612	0.0771	0.0807	0.0058	94.2
Gender	0.1607	0.1155	−0.0452	0.0048	0.0068	0.0004	89.6
Age	0.0217	0.0148	−0.0069	0.0000	0.0001	0.0000	67.4
lghrs	0.2878	0.3561	0.0683	0.0025	0.0071	0.0004	72.6
π_{1639}	0.0503	0.0586	0.0082	0.0000	0.0001	0.0000	—
π_{1050}	0.1199	0.1235	0.0037	0.0001	0.0002	0.0000	—
π_{2690}	0.1479	0.1336	−0.0143	0.0004	0.0006	0.0000	—
Method: *adj*-1							
Intercept	−5.5848	−7.9952	−2.4105	0.5128	6.3222	0.1760	22.0
Gender	0.1607	0.2297	0.0690	0.0203	0.0251	0.0021	96.8
Age	0.0217	0.0337	0.0119	0.0001	0.0002	0.0000	81.4
lghrs	0.2878	0.5645	0.2767	0.0139	0.0905	0.0034	41.4
π_{1639}	0.0503	0.0448	−0.0055	0.0001	0.0001	0.0000	—
π_{1050}	0.1199	0.1776	0.0577	0.0005	0.0038	0.0001	—
π_{2690}	0.1479	0.2310	0.0831	0.0029	0.0098	0.0005	—
Method: *adj*-2							
Intercept	−5.5848	−5.5502	0.0346	0.1393	0.1402	0.0085	93.2
Gender	0.1607	0.2258	0.0652	0.0124	0.0166	0.0012	94.4
Age	0.0217	0.0292	0.0075	0.0000	0.0001	0.0000	86.2
lghrs	0.2878	0.2275	−0.0603	0.0037	0.0074	0.0004	79.8
π_{1639}	0.0503	0.0454	−0.0049	0.0001	0.0001	0.0000	—
π_{1050}	0.1199	0.1269	0.0071	0.0003	0.0004	0.0000	—
π_{2690}	0.1479	0.1811	0.0332	0.0014	0.0025	0.0002	—

The foregoing results motivate the question of what might be an appropriate number of cases (or controls, whichever is the smaller) for the asymptotic results to hold and the adj-2 method's performance to be adequate. Peduzzi *et al.* [8] suggest a rule of 10 cases to each covariate in the context of logistic regression, but we fear that such a general recommendation is not possible in our setting, due to the complications provided by the multiple correlations among covariates and the potential for misclassification. Still, we attempt to make a more general recommendation for settings like ours by repeating the work in Table IV for a larger

Table V. Simulation results with the null case and sample size $n = 1000$, based on 200 simulated data sets.

Covariate (j)	True$_j$	Mean$_j$	Bias$_j$	Variance$_j$	MSE$_j$	$\widehat{se}$(MSE$_j$)	CP$_j$ (per cent)
Method: *standard*							
Intercept	−3.27	−3.3895	−0.1195	0.1723	0.1857	0.0185	95.5
Gender	0.00	0.0585	0.0585	0.1424	0.1452	0.0172	94.5
Age	0.00	0.0021	0.0021	0.0001	0.0001	0.0000	96.0
lghrs	0.00	−0.0032	−0.0032	0.0071	0.0071	0.0007	95.5
Method: *naive*							
Intercept	−3.27	−3.9311	−0.6611	0.2454	0.6812	0.0605	71.0
Gender	0.00	0.1215	0.1215	0.1306	0.1447	0.0197	96.0
Age	0.00	0.0045	0.0045	0.0001	0.0002	0.0000	92.0
lghrs	0.00	0.1101	0.1101	0.0089	0.0210	0.0020	76.5
Method: *adj*-1							
Intercept	−3.27	−3.9825	−0.7125	1.7268	2.2251	0.4916	98.9
Gender	0.00	0.1183	0.1183	0.6075	0.6182	0.0922	100.0
Age	0.00	0.0016	0.0016	0.0008	0.0008	0.0001	98.9
lghrs	0.00	0.0314	0.0314	0.0580	0.0586	0.0134	96.8
Method: adj-2							
Intercept	−3.27	−2.8934	0.3766	0.2434	0.3840	0.0319	84.0
Gender	0.00	−0.0748	−0.0748	0.3702	0.3739	0.0752	96.5
Age	0.00	−0.0023	−0.0023	0.0003	0.0003	0.0000	97.0
lghrs	0.00	−0.0897	−0.0897	0.0112	0.0192	0.0018	83.5

sample size. Specifically, we resample (with replacement) $n = 4(3687) = 14\,748$ observations from the original 3687 observations. For each observation, we then generate its 'true' response value as before, obtaining roughly 540 cases ($Y = 1$). The results indicate that there is some dependence on the outcome of interest, with the adj-2 method typically having roughly equal MSE performance and slightly better coverage. For example, for the important lghrs covariate we obtain an MSE of 0.0057 for the naive method but 0.0051 for adj-2. Results for other, smaller sample sizes ($n = 2(3687)$, $3(3687)$) are similar but less dramatic. Overall, our results suggest the asymptotics support adjustment in this setting only when the cases of events exceeds 500.

Null case

It is known that with non-differential measurement errors, the naive method provides a fully efficient test of association (e.g. Reference [9]). However, with differential misclassification errors as in the case here, even under the null model, the naive method fails to draw correct inferences: its coverage probability can be much lower than the specified nominal level (Tables V, VI). In contrast, adj-2 works well. On the other hand, the naive method can still yield smaller MSEs of the estimates than that of adj-2.

In summary, if the goal is to draw statistical inference on parameters, adj-2 is better than the naive method. However, for the purpose of point estimation of parameters, with smaller sample sizes (e.g. 1000, which would result in only 37 events given our low event rate), the

Table VI. Simulation results with the null case and sample size $n = 3687$, based on 500 simulated data sets.

Covariate (j)	True$_j$	Mean$_j$	Bias$_j$	Variance$_j$	MSE$_j$	$\widehat{se}$(MSE$_j$)	CP$_j$ (per cent)
method: *standard*							
Intercept	−3.27	−3.3088	−0.0388	0.0417	0.0431	0.0032	95.0
Gender	0.00	0.0126	0.0126	0.0348	0.0349	0.0022	94.8
Age	0.00	0.0005	0.0005	0.0000	0.0000	0.0000	94.6
lghrs	0.00	0.0025	0.0025	0.0017	0.0017	0.0001	95.6
method: *naive*							
Intercept	−3.27	−3.8510	−0.5810	0.0515	0.3889	0.0131	22.6
Gender	0.00	0.1095	0.1095	0.0350	0.0470	0.0026	91.0
Age	0.00	0.0041	0.0041	0.0000	0.0000	0.0000	89.2
lghrs	0.00	0.1038	0.1038	0.0020	0.0128	0.0005	30.0
Method: *adj*-1							
Intercept	−3.27	−3.6170	−0.3470	0.1980	0.3180	0.0387	95.2
Gender	0.00	0.0221	0.0221	0.1405	0.1407	0.0117	97.2
Age	0.00	0.0006	0.0006	0.0001	0.0001	0.0000	96.6
lghrs	0.00	0.0089	0.0089	0.0078	0.0079	0.0017	96.6
Method: *adj*-2							
Intercept	−3.27	−2.7994	0.4706	0.0566	0.2779	0.0102	50.6
Gender	0.00	−0.0877	−0.0877	0.0688	0.0763	0.0053	95.4
Age	0.00	−0.0037	−0.0037	0.0001	0.0001	0.0000	94.6
lghrs	0.00	−0.0813	−0.0813	0.0026	0.0092	0.0004	63.4

value of adjusting for bias due to misclassification is unclear, due to the possible resulting increase in mean squared error of the associated parameter estimates. On the other hand, for larger sample sizes (e.g. 3687), the benefits of bias adjustment may well be more compelling (Table V).

Impact of misclassification error

The procedures we have compared ignore errors in the estimates for the sensitivity and specificity based on the validation study. To investigate the impact of misclassification error, we conducted a final, brief simulation study to investigate its consequence. The simulation set-up is similar to our earlier one for the non-null case except for a modification to the calculation of subject-specific specificity. By fitting a linear logistic regression model of the form in (2),

$$\text{logit}(\text{specificity}) = \alpha_0 + \alpha_1(\text{Resptype} = 1) + \alpha_2(\text{Resptype} = 2) + \alpha_3\,\text{lghrs} \tag{5}$$

to the validation study data, we obtain not only the point estimates of the regression coefficients $\widehat{\alpha} = (\widehat{\alpha}_0, \ldots, \widehat{\alpha}_3)'$, but also an asymptotic standard error estimate and multivariate normal distribution for $\widehat{\alpha}$. For each subject i then, we may draw a random sample of $\widehat{\alpha}$ from this asymptotic distribution and denote it as $\widehat{\alpha}^i$. Then the specificity for subject i can be calculated by plugging $\alpha = \widehat{\alpha}^i$ into equation (4). Repeating this process n times, we obtain a simulated data set. We then fit the model as before, using (2) directly and ignoring the variability of the resulting estimates.

Table VII. Simulation results with errors in validation estimates and sample size $n = 3687$, based on 500 simulated data sets.

Covariate (j)	True$_j$	Mean$_j$	Bias$_j$	Variance$_j$	MSE$_j$	$\widehat{se}$(MSE$_j$)	CP$_j$ (per cent)
Method: *standard*							
Intercept	−5.5848	−5.6484	−0.0637	0.1614	0.1652	0.0112	94.4
Gender	0.1607	0.1642	0.0035	0.0119	0.0118	0.0008	95.2
Age	0.0217	0.0218	0.0000	0.0000	0.0000	0.0000	96.8
lghrs	0.2878	0.2955	0.0077	0.0048	0.0048	0.0003	94.0
Method: *naive*							
Intercept	−5.5848	−5.3553	0.2294	0.1355	0.1879	0.0099	84.2
Gender	0.1607	0.1141	−0.0465	0.0105	0.0127	0.0007	90.0
Age	0.0217	0.0142	−0.0075	0.0000	0.0001	0.0000	76.6
lghrs	0.2878	0.3240	0.0363	0.0041	0.0054	0.0004	91.8
Method: *adj*-1							
Intercept	−5.5848	−8.1502	−2.5654	1.0343	7.6137	0.2881	40.9
Gender	0.1607	0.2277	0.0670	0.0496	0.0540	0.0070	98.8
Age	0.0217	0.0335	0.0117	0.0002	0.0003	0.0000	92.2
lghrs	0.2878	0.5859	0.2981	0.0263	0.1151	0.0056	76.7
Method: *adj*-2							
Intercept	−5.5848	−5.8822	−0.2974	0.4053	0.4930	0.0411	97.4
Gender	0.1607	0.2222	0.0615	0.0294	0.0332	0.0023	96.2
Age	0.0217	0.0303	0.0086	0.0001	0.0002	0.0000	92.4
lghrs	0.2878	0.2620	−0.0258	0.0102	0.0108	0.0007	93.4

The results for $n = 3687$ are summarized in Table VII. Unsurprisingly, because of the ignored variability in equation (2), all of the adjustment methods tend to give a lower confidence interval coverage than the nominal level. To account for the extra variability of the validation study estimates, a multiple imputation approach [10] should perhaps be adopted; this is a subject of current investigation.

ANALYSIS OF AGRICULTURAL INJURY DATA

We now apply both the naive and adj-2 methods to the data collected in RRIS-II and OATS (due to the presence of differential misclassification, we do not consider the adj-1 method). The results are presented in Table VIII. The first goal of the analysis is to detect whether there is any association between agricultural injury and any of the three risk factors: gender, age, and farm work time (lghrs). The significance levels obtained from the two methods are in general agreement with each other: at the usual 5 per cent significance level, both methods find a statistically significant association between the outcome and age and between the outcome and farm work time, whereas the association between the outcome and gender is not significant after adjusting for the other two risk factors. However, if we are interested in assessing the strength of association using the odds ratios and associated CIs, then we may prefer the adj-2 results, due to their generally superior performance in our previous simulation using the RRIS-II/OATS sample size, $n = 3687$. In addition, bias correction leads

Table VIII. Results from logistic regression with and without adjustment for (differential) misclassification for the agricultural injury data.

Covariate (j)	Method	$\hat{\beta}_j$	SE_j	95 per cent CI_j	P_j
Intercept	Naive	−5.5838	0.3594	(−6.2882, −4.8794)	<0.0001
Gender	Naive	0.1607	0.1065	(−0.0480, 0.3694)	0.1313
Age	Naive	0.0217	0.0067	(0.0086, 0.0348)	0.0012
lghrs	Naive	0.2876	0.0637	(0.1627, 0.4125)	<0.0001
Intercept	adj-2	−5.3132	0.44644	(−6.1882, −4.4382)	<0.0001
Gender	adj-2	0.2113	0.1649	(−0.1119, 0.5345)	0.2001
Age	adj-2	0.0299	0.0100	(0.0104, 0.0494)	0.0027
lghrs	adj-2	0.1893	0.0806	(0.0314, 0.3472)	0.0188

to estimates with more realistic (i.e. larger) standard errors that properly account for the differential misclassification in the data: Table VIII shows that all four of the 95 per cent CIs are wider under the adj-2 method.

DISCUSSION

In practice, the question of whether or not to adjust for possible misclassification of a binary outcome in logistic regression does not seem to have an easy or straightforward answer. Many factors, such as the misclassification rate, sample size, and the primary analytic question of interest, enter into the decision. The fundamental reason for this is the tradeoff between bias and variance, a general and well-known statistical property. Although an adjustment method can reduce the bias of a particular statistical estimate, it can at the same time introduce more variability, thus, yielding an estimator with a larger MSE than a naive, unadjusted method. Hence, if the primary analytic goal is to obtain point estimates of parameters (e.g. for the purpose of building a predictive model), an adjustment method may perform less well than an unadjusted one. This observation is not new; in fact, it tends to arise whenever a more complicated statistical method is used (see e.g. [5, Section 2.4]; [11, Appendix B.2]). However, as the sample size increases, in terms of yielding smaller MSEs, the performance of an adjustment method can often surpass that of an unadjusted method. Therefore, the 'gold standard' for making the decision in this case must be sensitivity analyses or pre-analytic simulation studies, in order to empirically evaluate the effects of adjusting or not adjusting for misclassification of an outcome. For example, a sensitivity analysis in our setting might redraw 3687 samples (with replacement) from our data set and recompute the parameters of interest under both the naive and our bias-corrected approaches. On the other hand, if the primary goal is interval estimation or hypothesis testing regarding the parameters, adjusting for possible misclassification is necessary. While some authors (e.g. Reference [12]) discourage the practice of interpreting only point estimates, we would argue that adjustment may be suboptimal if point estimation is the main goal; it need not be the *only* goal for our findings to be noteworthy.

Finally, it is well known that non-differential misclassification of an outcome biases the association parameters toward zero (the null value) when this is not taken into account [13–16].

However, the direction of this bias can change if there is differential misclassification. In the agricultural injury study, log-working time is negatively associated with the specificity; as a result, its association estimate is biased *away* from zero (see Table VI). Our simulation results show that the adjustment method under the assumption of nondifferential misclassification (adj-1) not only does not reduce, but actually introduces extra bias when the misclassification is in fact differential, for which case adj-2 performs much better. Hence, it appears important to investigate any possible association between the sensitivity or specificity and other subject characteristics before deciding on any adjustment method.

ACKNOWLEDGEMENTS

The authors thank Colleen Renier, and two reviewers for many helpful suggestions and comments. The work of the second author was supported in part by an NIH grant R01-HL65462. The work of the fourth author was supported in part by NIAID grant R01-AI41966 and NSF/EPA grant SES 99-78238.

REFERENCES

1. Rothman KJ, Greenland S. *Modern Epidemiology* (2nd edn). Williams & Wilkins: Philadelphia, Lippincott, 1998.
2. Gerberich SG, Gibson RW, Gunderson PD, Melton III LJ, French LR, Renier C, Erdman AG, True JA, Carr P, Elkington J. Validity of trauma reporting in the agricultural community. *Journal of Occupational Accidents* 1990; **12**:200–201.
3. Gerberich SG, Gibson RW, Gunderson PD, French LR, Melton LJ III, Erdman A, Smith P, True JA, Carr WP, Elkington J, Renier CM, Andreasson LR. The olmsted agricultural trauma study (OATS): a population-based effort. *Report to the Centers for Disease Control*, Regional Injury Prevention Research Center, Minneapolis, 1991 (NTIS # PB92-107168/AS).
4. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainy. *American Journal of Epidemiology* 1997; **146**(2):195–203.
5. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models.* Chapman & Hall: London, U.K., 1995.
6. Gerberich SG. Etiology and consequence of injuries among children in farm households: regional rural injury study-II. *Grant Application Report*, 1997.
7. Hastie T, Tibshirani R. *Generalized Additive Models.* Chapman & Hall: New York, NY, 1990.
8. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
9. Tosteson TD, Tsiatis AA. The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika* 1988; **75**:507–514.
10. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* Wiley: New York, NY, 1987.
11. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis* (2nd edn). CRC Press, Chapman & Hall: Boca Raton, FL, 2000.
12. Simon R. Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine* 1986; **105**:429–435.
13. Barron BA. The effects of misclassification on estimation of relative risk. *Biometrics* 1977; **33**:414–418.
14. Copeland KT, Checkoway H, McMichael AJ *et al*. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology* 1977; **105**:488–495.
15. Fleiss JL. *Statistical Methods for Rates and Proportions.* Wiley: New York, 1981.
16. Kleinbaum DG, Kupper L, Morgenstern H. *Epidemiologic Research*: *Principles and Quantitative Methods.* Lifetime Learning Inc.: San Francisco, CA, 1982.

