American Journal of Epidemiology
Copyright © 1997 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 146, No. 2
Printed in U.S.A.

Logistic Regression When the Outcome Is Measured with Uncertainty

Laurence S. Magder[1] and James P. Hughes[2]

In epidemiologic research, logistic regression is often used to estimate the odds of some outcome of interest as a function of predictors. However, in some datasets, the outcome of interest is measured with imperfect sensitivity and specificity. It is well known that the misclassification induced by such an imperfect diagnostic test will lead to biased estimates of the odds ratios and their variances. In this paper, the authors show that when the sensitivity and specificity of a diagnostic test are known, it is straightforward to incorporate this information into the fitting of logistic regression models. An EM algorithm that produces unbiased estimates of the odds ratios and their variances is described. The resulting odds ratio estimates tend to be farther from the null but have greater variance than estimates found by ignoring the imperfections of the test. The method can be extended to the situation where the sensitivity and specificity differ for different study subjects, i.e., nondifferential misclassification. The method is useful even when the sensitivity and specificity are not known, as a way to see the degree to which various assumptions about sensitivity and specificity affect one's estimates. The method can also be used to estimate sensitivity and specificity under certain assumptions or when a validation subsample is available. Several examples are provided to compare the results of this method with those obtained by standard logistic regression. A SAS macro that implements the method is available on the World Wide Web at http://som1.ab.umd.edu/Epidemiology/software.html. *Am J Epidemiol* 1997;146:195–203.

biometry; diagnosis; epidemiologic methods; logistic models; sensitivity and specificity

Epidemiologists often use logistic regression to estimate the effect of various predictors on some binary outcome of interest. In brief, this involves assuming that the log of the odds of the outcome is a linear function of the predictors and estimating the coefficients of the function using maximum likelihood.

This method requires data that consist of known values for the binary outcome and the predictors. However, in some research, the outcome of interest is not measured perfectly. For example, Magder et al. (1) used logistic regression to estimate the degree to which various factors predicted *Chlamydia trachomatis* infection in a sexually transmitted diseases clinic population. The presence or absence of chlamydia infection was determined by tissue culture. Unfortunately, the sensitivity of culture for detecting chlamydia infection was believed to be much less than 100 percent. Thus it is probable that many of the subjects assumed to be negative in the analysis were misclassified. This could lead to biased estimates (2) and incorrect inferences (3).

One approach to this problem is to report the biased estimates with the caveat that if the misclassification is nondifferential, estimates of odds ratios will be biased toward the null. However, another approach is possible when the sensitivity and specificity of the test are known or can be assumed to take certain values. In that case, unbiased estimates of the odds ratios can be found by incorporating the values of sensitivity and specificity into the likelihood and using maximum likelihood estimation. A number of authors (e.g., 4–7) have described how this might be implemented in the framework of log-linear models for cross-classified data. One drawback to log-linear models is that unlike logistic regression models, they cannot be used to model the effect of a continuous predictor without breaking the predictor into categories.

In this paper, we show that it is relatively easy to incorporate information on the values of sensitivity and specificity into the estimation of the parameters in a logistic regression model. This can be done by modifying software used for fitting standard logistic regression models with the use of an EM algorithm (8). The resulting procedure has an interesting and intuitive interpretation, as described below. There is

Received for publication July 10, 1996, and accepted for publication March 5, 1997.

Abbreviations: ELISA, enzyme-linked immunosorbent assay; HIV, human immunodeficiency virus; MLE, maximum likelihood estimate.

[1] Department of Epidemiology and Preventive Medicine, University of Maryland at Baltimore, Baltimore, MD.

[2] Department of Biostatistics, University of Washington, Seattle, WA.

no requirement for the sensitivity and specificity to be the same for all observations; thus the method can accommodate differential misclassification. If the sensitivity and/or specificity of the classification procedure are not known, the method can be useful in showing the degree to which estimates change under various assumptions about sensitivity and specificity. The method can also be used to estimate sensitivity and specificity under various assumptions or when a validation subsample is available.

First, we describe briefly how information about sensitivity and specificity can be incorporated into the estimation of odds ratios from two-by-two tables. Next, we describe how to incorporate information about sensitivity and specificity into the fitting of logistic regression models. We then give some examples and finally, provide a few additional comments.

ESTIMATION FROM TWO-BY-TWO TABLES

It is informative to first consider the form that maximum likelihood estimates (MLEs) of odds ratios take when there is only one dichotomous predictor and the sensitivity and specificity of the outcome are known and assumed to be the same for each observation. For simplicity, the outcome is referred to as "disease" (D) and the predictor as "exposure" (E).

As usual, let a, b, c, and d stand for the number of observations that are classified into each predictor/outcome category (table 1). In addition, let $\hat{p}_{T+|E} = a/(a+b)$ and $\hat{p}_{T+|E} = c/(c+d)$ denote the standard estimates of the probability of disease given exposure and nonexposure, respectively. The standard estimate for the odds ratio can be written as

$$\widehat{OR}_{standard} = \frac{\hat{p}_{T+|E}}{1-\hat{p}_{T+|E}} \div \frac{\hat{p}_{T+|\bar{E}}}{1-\hat{p}_{T+|\bar{E}}} = \frac{ad}{bc}$$

This is, in fact, the MLE of the odds ratio under the assumption that the testing procedure used to diagnose disease is perfectly accurate. If the procedure is imperfect, the values a, b, c, and d will reflect misclassifications, and the estimates above will be biased.

Now let *sens* stand for the sensitivity of the procedure used to classify the outcome, i.e.,

sens = probability that a study subject is classified as diseased given that the subject is truly diseased.

TABLE 1. Notation for cell counts in two-by-two tables

	Classified as	
	Diseased	Not diseased
Exposed	a	b
Not exposed	c	d

Similarly, let *spec* stand for the specificity of the procedure, i.e.,

spec = probability that a study subject is classified as not diseased given that the subject is truly not diseased.

If the values of *sens* and *spec* are known, then the MLE of the odds ratio can be shown (see Appendix) to take the following form:

$$\widehat{OR} = \frac{\hat{p}_{T+|E} - (1 - spec)}{sens - \hat{p}_{T+|E}} \div \frac{\hat{p}_{T+|\bar{E}} - (1 - spec)}{sens - \hat{p}_{T+|\bar{E}}}. \quad (1)$$

This formula can be used only if all the numerators and denominators in the formula are positive. If one of these is negative, the MLE is 0 or infinity. To understand why, consider the numerator of the first fraction. If $\hat{p}_{T+|E}$ is less than $(1 - spec)$, then the dataset would contain fewer exposed patients classified as diseased than would be expected if all the exposed were truly disease free and all those classified as diseased were false-positives. Thus the data would be most consistent with the possibility that there is no risk of disease in the exposed, and the correct MLE would be 0. Similarly, if $\hat{p}_{T+|E}$ is greater than *sens*, then the data would be most consistent with the possibility that all the exposed were diseased, and the correct MLE for the odds ratio would be infinity.

Assuming that $\widehat{OR}$ is not equal to 0 or infinity, it will always be farther from 1 than $\widehat{OR}_{standard}$. This can be seen as follows: Assume that $\hat{p}_{T+|E} > \hat{p}_{T+\bar{E}}$ so that $\widehat{OR}_{standard} > 1$. Then, when sensitivity and specificity are not equal to 1, we can write

$$\widehat{OR} = \frac{\hat{p}_{T+|E} - (1 - spec)}{\hat{p}_{T+|\bar{E}} - (1 - spec)} \times \frac{sens - \hat{p}_{T+|\bar{E}}}{sens - \hat{p}_{T+|E}}$$

$$> \frac{\hat{p}_{T+|E}}{\hat{p}_{T+|\bar{E}}} \times \frac{sens - \hat{p}_{T+|\bar{E}}}{sens - \hat{p}_{T+|E}}$$

$$> \frac{\hat{p}_{T+|E}}{\hat{p}_{T+|\bar{E}}} \times \frac{1 - \hat{p}_{T+|\bar{E}}}{1 - \hat{p}_{T+|E}}$$

$$= \widehat{OR}_{standard}.$$

A similar argument can be made if $\widehat{OR}_{standard} < 1$. Thus incorporating the information about sensitivity and specificity in the estimation process compensates for the bias of the standard estimate toward the null.

ESTIMATING PARAMETERS IN LOGISTIC REGRESSION MODELS

Using known (or assumed) values of sensitivity and specificity

The standard approach to the estimation of parameters in a logistic regression model is to use maximum likelihood. Any algorithm designed to do this can be modified to incorporate information about the sensitivity and specificity of the outcome measurement through the use of an EM algorithm.

The resulting procedure can be described in an intuitive way. Essentially, the procedure is to perform standard logistic regression considering each study subject as both diseased and not diseased with weights determined by the probability that the study subject is truly diseased given the data. For example, suppose a study subject has been classified as diseased by the imperfect diagnostic test. Further suppose that given the sensitivity and specificity of the test and the values of that subject's covariates, it is calculated that there is 90 percent probability that the subject is truly diseased. Then standard logistic regression would be performed with that subject entered twice: once as diseased with a weight of 0.9 and once as undiseased with a weight of 0.1. Since the probability that a subject is truly diseased depends in part on the value of the logistic regression parameters, these probabilities need to be recalculated after the logistic regression parameters are estimated. This leads to new probabilities, which in turn lead to new estimates of the regression parameters. The processes of estimating the probabilities (the "E" step of the EM algorithm) and the logistic regression parameters (the "M" step) are repeated alternately until the parameter estimates stop changing.

To explain the algorithm more formally, let $Y_i = 1$ if the ith individual is truly diseased, and 0 otherwise; $T_i = 1$ if the ith individual is classified as diseased, and 0 otherwise, for i from 1 to n. Given some known covariates, $X_{1i}, X_{2i}, \ldots X_{ki}$, the logistic regression model entails the assumption that

$$Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})} \quad (2)$$

for some unknown parameters, $\beta_0, \beta_1, \beta_2, \ldots \beta_k$.

Now, let $\hat{Y}_i$ be the probability that the ith individual is truly diseased, given the values of T_i and $X_{1i}, X_{2i}, \ldots X_{ki}$. If $T_i = 1$ (i.e., the ith individual is classified as diseased), then $\hat{Y}_i$ is the predicted value of a positive test for the ith individual. By Bayes' theorem this is

$$\hat{Y}_i = \frac{Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)sens}{Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)sens + Prob(Y_i = 0|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)(1 - spec)}.$$

Similarly, if $T_i = 0$, then

$$\hat{Y}_i = \frac{Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)(1 - sens)}{Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)(1 - sens) + Prob(Y_i = 0|X_{1i}, \ldots X_{ki}, \beta_0, \ldots \beta_k)spec}.$$

To find maximum likelihood estimates for the βs, one can proceed as follows: From some initial estimates of the βs, the $\hat{Y}_i$s are calculated. Standard logistic regression is then performed with each subject included as both diseased and undiseased, with weights equal to $\hat{Y}_i$ and $(1 - \hat{Y}_i)$, respectively. This leads to updated estimates of the βs, which in turn leads to new values for the $\hat{Y}_i$s. The process is repeated until the estimates converge. For a justification of the fact that this algorithm will lead to maximum likelihood estimates, see appendix 2.

It is straightforward to incorporate nondifferential misclassification into the fitting algorithm. The sensitivity and specificity can be given different values for different individuals, depending on the values of their covariates.

Standard errors for the parameter estimates can be approximated in a standard way using the inverse of the information matrix. This matrix can be expressed in a relatively simple expression and is given in appendix 3.

If sensitivity and specificity are unknown

If the sensitivity and specificity of the tests are not known, it may be formally possible to estimate them using maximum likelihood. As shown in the appendix, sensitivity and specificity can be estimated simultaneously with the βs by setting them equal to

$$\widehat{sens} = \frac{\sum_{(i:T_i=1)} \hat{Y}_i}{\sum_{(\text{all } i)} \hat{Y}_i} \quad (3)$$

and

$$\widehat{spec} = \frac{\sum_{(i:T_i=0)} (1 - \hat{Y}_i)}{\sum_{(\text{all } i)} (1 - \hat{Y}_i)} \quad (4)$$

at each iteration of the EM algorithm.

There is one major problem with this approach, however. In the simplest situation (i.e., two-by-two tables) and in saturated models, the parameters in the model (*sens*, *spec*, $\beta_0, \ldots \beta_k$) are not identifiable. In other words, different parameter sets lead to the same probability distribution, and therefore there are an infinite number of maximum likelihood estimates. This problem does not occur if the model incorporates some smoothing assumptions (such as the assumption that the log odds of disease is a linear function of some continuous predictor or the assumption that there is no interaction between categorical predictors). In that case, the parameters are identifiable and can be estimated. However, it is clear that any information about sensitivity and specificity in the data would be the result of the smoothing assumptions, which are often difficult to verify and are generally only approximately true. Thus it might be inadvisable to estimate sensitivity and specificity in this manner unless there is strong reason to believe in some of the smoothing assumptions.

In some research, when the sensitivity and specificity of the diagnostic test are not known, a "gold standard" test is administered in addition to the fallible test on a subsample of the study subjects. This can be referred to as a "validation" subsample. In this case, for those in the validation subsample, $\hat{Y}_i$ can be given the value determined by the gold standard test, and formulas 3 and 4 can be used to estimate sensitivity and specificity, respectively. Note that the estimates of sensitivity and specificity would still be based on all the data, not just the validation subsample.

EXAMPLES

Example 1: Predictors of *Chlamydia trachomatis* infection

Several thousand patients attending a sexually transmitted diseases clinic in Denver, Colorado, from 1981 to 1983 were cultured for *C. trachomatis* infection to determine what patient characteristics might predict chlamydia infection (1). As noted in that paper, the sensitivity of chlamydia culture was believed to be much less than 100 percent. Here some of those data are reanalyzed to determine how the results might change if the sensitivity of chlamydia culture were assumed to be 70 percent.

Specifically, a model was fitted to determine whether the appearance or amount of discharge is predictive of chlamydia infection in men with discharge. The model was fitted twice: once with a standard logistic regression package (PROC LOGISTIC) (9) and once using an EM algorithm incorporating the assumptions that chlamydia culture is 100 percent specific but only 70 percent sensitive. The dataset consisted of 1,190 men of whom 281 (24 percent) had a positive culture for chlamydia.

The results are presented in table 2. Notice that every odds ratio estimate from the fit that incorporated the imperfect sensitivity is farther from 1 than the corresponding estimate from standard logistic regression. This is consistent with the observation from the two-by-two tables and the fact that nondifferential misclassification makes standard estimates biased toward the null. Notice also that the standard error of the estimates are all greater than those found using the standard method. This is reflective of the fact that imperfect sensitivity results in a loss of information.

Perhaps most striking is the fact that the estimates change so little under the assumption of 70 percent sensitivity. This is due to the fact that the standard estimate of the odds ratio is only slightly biased when the probability of the outcome is low, and the specificity is 100 percent. This can be confirmed by trying some values in equation 1. This finding is related to the fact that the standard estimate of the risk ratio is fully unbiased, no matter what the sensitivity of the test, as long as the specificity is 100 percent.

Example 2: Predictors of smoking cessation

Sexton and Hebel (10) reported the results of a randomized trial of a smoking cessation program among pregnant women. In this example, only those women who were randomized to receive the program are considered. Among that group, it is of interest to know what patient characteristics predict successful smoking cessation. This outcome, smoking cessation, was measured by self-report; and there was some concern about the accuracy of the patients' reports. The researchers believed that among those who actually quit smoking, the probability that they reported quitting is quit high (sensitivity approximately 100 percent) (R. Hebel, University of Maryland at Baltimore, personal communication, 1996). However, the researchers were concerned that among those who did

TABLE 2. Logistic regression estimates of association between discharge appearance and chlamydia infection among men presenting with discharge, Denver Metro Health Clinic for Sexually Transmitted Diseases, September 1981 to June 1983

Variable	*β* estimate	Standard error	Odds ratio estimate	95% confidence interval
		Method 1: Standard logistic regression		
Age (years)				
<20	0.29	0.32	1.34	0.72–2.49
20–29	0.48	0.15	1.61	1.19–2.17
≥30	0.00	Reference		
Discharge color				
Clear	0.00	Reference		
White	0.28	0.17	1.32	0.95–1.85
Yellow	−0.26	0.24	0.77	0.49–1.23
Discharge amount				
Scant	0.00	Reference		
Moderate	0.23	0.17	1.25	0.91–1.73
Profuse	−0.51	0.26	0.60	0.36–1.01
		Method 2: Assuming chlamydia culture is 100% specific and 70% sensitive		
Age (years)				
<20	0.34	0.37	1.40	0.68–2.87
20–29	0.56	0.18	1.74	1.24–2.46
≥30	0.00	Reference		
Discharge color				
Clear	0.00	Reference		
White	0.35	0.20	1.42	0.96–2.11
Yellow	−0.29	0.27	0.74	0.44–1.26
Discharge amount				
Scant	0.00	Reference		
Moderate	0.26	0.20	1.29	0.88–1.91
Profuse	−0.57	0.29	0.57	0.32–1.00

not quit, some might report that they had quit (specificity < 100 percent).

To determine what impact a specificity of 90 percent would have on estimates of the degree to which various characteristics predicted smoking cessation, we used the methods described in this paper. The strongest predictor of smoking cessation was the number of cigarettes smoked per day at the time of randomization. Among those who reported smoking less than one pack per day at that time, 101/254 (40 percent) reported quitting whereas among those who were smoking a pack or more, only 15/107 (14 percent) reported quitting. The standard estimate of the odds ratio comparing those who smoke less than one pack with those who smoke more is 4.0. However, this is likely to be an underestimate of the odds ratio if the self-report is only 90 percent specific. In that case, the estimate that 14 percent of the heavier smokers quit is likely to be much too high (because even if none of them quit, we would expect to see about 10 percent reporting quitting). Using equation 1, the estimate for the odds ratio was found to be 10.6. The results of this bivariate analysis do not change much after age and education are controlled for in a logistic regression analysis (table 3).

ADDITIONAL COMMENTS

Applicability to case-control studies

In the classic case-control study, a sample of cases (i.e., those diagnosed with some disease of interest) is compared with a sample of controls (those without the disease) with respect to history of exposures and other possible predictors. A typical way of analyzing such studies is to ignore the fact that the number of cases and controls was fixed by design and to perform logistic regression using case-control status as the binary outcome. This is justified by the fact that the odds ratios will be the same whatever the actual sampling fractions of the cases and controls (11).

Despite the fact that standard logistic regression can be applied to case-control studies in this manner, the methods described in this paper cannot. The reason is apparent from the fitting algorithm. Recall that the fitting algorithm is based on calculating the probability that each subject truly has disease given the data. This

TABLE 3. Logistic regression estimates of association between patient characteristics and smoking cessation among pregnant women participating in a smoking cessation program, Baltimore, Maryland, metropolitan area, 1978–1981

Variable	β estimate	Standard error	Odds ratio estimate	95% confidence interval
		Method 1: Standard logistic regression		
Age (years)				
<20	0.00		Reference	
20–29	−0.47	0.37	0.62	0.30–1.29
≥30	−0.18	0.45	0.83	0.35–2.00
Education				
<High school	0.00		Reference	
High school graduate	0.10	0.31	1.11	0.60–2.04
Some college	−0.42	0.38	0.66	0.31–1.37
Previous smoking history (pack/day)				
≥1	0.00		Reference	
<1	1.46	0.31	4.31	2.34–7.96
		Method 2: Assuming 90% specificity and 100% sensitivity		
Age (years)				
<20	0.00		Reference	
20–29	−0.55	0.47	0.57	0.23–1.45
≥30	−0.20	0.58	0.82	0.26–2.56
Education				
<High school	0.00		Reference	
High school graduate	−0.05	0.42	0.95	0.42–2.18
Some college	−0.75	0.53	0.47	0.17–1.33
Previous smoking history (pack/day)				
≥1	0.00		Reference	
<1	2.57	0.99	13.11	1.87–91.80

in turn is calculated from the sensitivity and specificity of the test and an estimate of the underlying probability of disease. In case-control studies, there is no way to estimate the underlying probability of disease since the numbers of cases and controls are fixed by design. In other words, if case-control status is determined by a fallible diagnostic test, there is no way to estimate what proportion of the cases are truly cases and what proportion of controls are truly controls. This point was previously noted by Chen (7).

An alternative approach to the application of logistic regression to case-control data, as described by Schlesselman (12), is to model the presence or absence of a particular exposure as the binary outcome in the logistic model. This also leads to estimates of the odds ratio relating that particular exposure to disease, controlling for the other variables in the model. If that approach is taken, the method described in this paper can be used with case-control data to take into account uncertainty in exposure classification.

Taking advantage of all the information available regarding the outcome

In many studies, more information than just a single test result is available regarding whether a subject has a disease (13). This is the case when patients are classified as diseased or not diseased based on the value of some continuous measure. For example, the diagnosis of diabetes might be made based on the value of a blood glucose measurement. Patients whose blood glucose exceeds some cut point would be classified as diabetic. In this example, the actual blood glucose measure contains more information about whether the patient is diabetic than the dichotomous test result. More information is also available when a diagnosis is based on a series of diagnostic tests. For example, human immunodeficiency virus (HIV) is often diagnosed based on the results of an enzyme-linked immunosorbent assay (ELISA) test followed by a confirmatory Western blot test. Those positive on both tests are classified as HIV positive, while those negative on one or the other test are classified as negative. This classification clearly contains less information than the individual test results. The probability that a subject has HIV might be different for one who is ELISA positive/Western blot negative than one who is ELISA negative and had no Western blot test done despite the fact that both would be classified as disease free.

This reduction of diagnostic information into a dichotomous classification is sometimes necessary in clinical decision-making. However, it entails a loss of information, which is not necessary in a research setting. The approach described in this paper can be used to take advantage of all the diagnostic information. Let Z_i stand for all the diagnostic information available. Then this information can be incorporated into the estimation process by estimating the probability that the ith subject has disease, $\hat{Y}_i$, as

$$\frac{Prob(Y_i = 1|X_{1i}, \dots X_{ki}, \beta_0, \dots \beta_k) Prob(Z_i|Y_i = 1)}{Prob(Y_i = 1|X_{1i}, \dots X_{ki}, \beta_0, \dots \beta_k) Prob(Z_i|Y_i = 1) + Prob(Y_i = 0|X_{1i}, \dots X_{ki}, \beta_0, \dots \beta_k) Prob(Z_i|Y_i = 0)}$$

at the E step of the EM algorithm. This requires knowledge of the value of $Prob(Z_i|Y_i = 1)$ and $Prob(Z_i|Y_i = 0)$. These values are known for some diagnostic tests based on continuous measures. In fact, some authors have recommended that these probabilities be used to make decisions about a patient even in a clinical setting (14).

Application to other forms of binary regression

The E step of the algorithm does not require the assumption that the log odds of disease is a linear function of the predictors. The function $Prob(Y_i = 1|X_{1i}, \dots, X_{ki}, \beta_0, \dots \beta_k)$ can take any form. Thus algorithms that fit other binary regression models, such as nonlinear models, can be modified as described in this paper to allow for imperfect sensitivity and specificity.

Estimates diverging to infinity

Estimates of logistic regression parameters with the method described in this paper are somewhat more likely to diverge to positive or negative infinity than those of standard logistic regression. This would occur in situations comparable to those that make one of the denominators or numerators of expression 1 negative. For example, if the logistic regression model includes a categorical predictor and if the proportion of subjects classified as diseased for one of the categories is less than ($1 - spec$), then the data would be most consistent with the possibility that there was no disease in that subgroup. In that case, the parameter estimate would diverge to negative infinity and the odds ratio estimate would be 0. Since this estimate is on the boundary of the likelihood surface, standard likelihood-based methods for inference about that parameter could not be used.

When predictor variables are measured with uncertainty

The methods described in this paper do not apply to the situation when predictor variables are measured with uncertainty. That situation leads to a more difficult problem because it requires the additional specification of a distribution for the predictor variables. A number of researchers have extended logistic regression to that situation (see Rosner et al. (15) and the references therein).

Availability of software

A SAS (9) macro has been written to implement the methods described in this paper. It can be found on the World Wide Web at http://som1.ab.umd.edu/Epidemiology/software.html.

ACKNOWLEDGMENT

The authors thank Dr. Franklyn N. Judson, for permission to use the data from the Denver Metro Health Clinic, and Drs. Mary Sexton and Richard Hebel, for permission to use the data on the smoking cessation study.

REFERENCES

1. Magder LS, Harrison HR, Ehret JM, et al. Factors related to genital *Chlamydia trachomatis* and its diagnosis by culture in a sexually transmitted disease clinic. Am J Epidemiol 1988; 128:298–308.
2. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 1977;105:488–95.
3. Quade D, Lachenbruch PA, Whaley FS, et al. Effects of misclassifications on statistical inferences in epidemiology. Am J Epidemiol 1980;111:503–15.
4. Chen TT. Log-linear models for categorical data with misclassification and double sampling. J Am Stat Assoc 1979;74: 481–8.
5. Espeland MA, Hui SL. The consultants forum: a general approach to analyzing epidemiologic data that contain misclassification errors. Biometrics 1987;43:1001–12.
6. Hochberg Y. On the use of double sampling schemes in analyzing categorical data with misclassification errors. J Am Stat Assoc 1977;72:914–21.
7. Chen TT. A review of methods for misclassified categorical data in epidemiology. Stat Med 1989;8:1095–106.
8. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 1977;39:1–38.
9. SAS Institute, Inc. PROC LOGISTIC, version 6.10. Cary, NC: SAS Institute, Inc., 1996.
10. Sexton M, Hebel JR. A clinical trial of change in maternal smoking and its effect on birth weight. JAMA 1984;251: 911–15.
11. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. (IARC scientific publication no. 32). Lyon: International Agency for Research

on Cancer, 1980:202–5.

12. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York, NY: Oxford University Press, 1982:267–9.
13. Jones RH, McClatchey MW. Beyond sensitivity, specificity and statistical independence. Stat Med 1988;7:1289–95.
14. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients. JAMA 1994;271:9–13.
15. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990;132:734–45.
16. Louis TA. Finding the observed information matrix when using the EM algorithm. J R Stat Society B 1982;44:226–33.

APPENDIX 1

Derivation of expression 1

Let $p_{D|E}$ stand for the probability of disease given exposure, and $p_{T+|E}$ stand for the probability of testing positive given exposure. Then $p_{T+|E} = p_{D|E} sens + (1 - p_{D|E})(1 - spec)$. This implies

$$p_{D|E} = \frac{p_{T+|E} - (1 - spec)}{sens - (1 - spec)}.$$

Similar definitions and arguments give

$$p_{D|\bar{E}} = \frac{p_{T+|\bar{E}} - (1 - spec)}{sens - (1 - spec)}.$$

This implies that the ratio of the odds of disease between exposed and unexposed is

$$OR = \frac{p_{T+|E} - (1 - spec)}{sens - p_{T+|E}} \div \frac{p_{T+|\bar{E}} - (1 - spec)}{sens - p_{T+|\bar{E}}}.$$

Replacing $p_{T+|E}$ and $p_{T+|\bar{E}}$ in this expression with their MLEs results in the MLE for the odds ratio.

APPENDIX 2

More detailed description of the EM algorithm

Here we describe the algorithm for the case when sensitivity and specificity are not known. When they are known, the algorithm is virtually the same, without the need to estimate sensitivity and specificity at each step.

If an individual tests positive ($T_i = 1$), the result could be either a false positive or a true positive. A similar statement can be made about a negative test. This leads to the following likelihood function:

$$L(\beta_0, \ldots, \beta_k, sens, spec|T_1, T_2, \ldots T_n) =$$

$$\prod_{i=1}^{n}(prob\,(Y_i = 1)sens + prob(Y_i = 0)(1 - spec))^{T_i}$$

$$(prob\,(Y_i = 1)(1 - sens) + prob(Y_i = 0)spec)^{1-T_i}, \quad (5)$$

where $Prob(Y_i = 1)$ and $Prob(Y_i = 0)$ are functions of the β as defined in expression 2. The objective is to find the values *sens*, *spec*, and β, which maximize this expression.

If we had observed the values of Y_i, then the probability of the data would take a simpler form:

$$L(\beta_0, \ldots, \beta_k, sens, spec|Y_1, Y_2, \ldots Y_n) =$$

$$\prod_{i=1}^{n}(prob\,(Y_i = 1))^{Y_i}(prob(Y_i = 0))^{1-Y_i}$$

$$\prod_{i=1}^{n}(sens^{T_i}\,(1 - sens)^{(1-T_i)})^{Y_i}\prod_{i=1}^{n}(spec^{(1-T_i)}$$

$$(1 - spec)^{T_i})^{1-Y_i}. \quad (6)$$

This is the likelihood for standard logistic regression multiplied by two likelihoods that would result from binomial distributions.

Expression 5 can be referred to as the "incomplete data likelihood" and expression 6 as the "complete data likelihood." Dempster et al. (8) showed that the incomplete data likelihood can be maximized by iteratively maximizing the expected value of the complete data log likelihood, where the expectation is taken over the distribution of the complete data given the incomplete data, the current values of the βs, and the sensitivity and specificity of the test. Since in this case the complete data log likelihood is linear with respect to the Y_is, this can be done by maximizing the complete data log likelihood after each unknown Y_i is replaced with the expected value of Y_i given T_i, sensitivity, specificity, and the current values of the βs. This leads to the algorithm described in the body of the paper.

APPENDIX 3

Estimating the variance of the logistic regression parameter estimates

The variance/covariance matrix of the parameter estimates is approximated by the inverse of the matrix representing the negative of the second derivative of the log likelihood function evaluated at the maximum likelihood estimates. This matrix is referred to as the observed information matrix, I_{obs}. Let

$$I_{obs} = \sum_i I_{obs,i},$$

where $I_{obs,i}$ is the contribution of the ith subject to the information matrix. For standard logistic regression,

$$I_{obs,i} = X_i X_i^T(\hat{p}_i(1 - \hat{p}_i)),$$

where X_i is the vector of covariates for the ith subject, and

$$\hat{p}_i = Prob(Y_i = 1|X_{1i}, \ldots X_{ki}, \hat{\beta}_0, \ldots \hat{\beta}_k).$$

Using the method described by Louis (16), it can be shown that for logistic regression when the outcome is measured with uncertainty,

$$I_{obs,i} = X_i X_i^T(\hat{p}_i(1 - \hat{p}_i) - \hat{Y}_i(1 - \hat{Y}_i)).$$