Published in final edited form as:

J Stat Plan Inference. 2013 November 1; 143(11): . doi:10.1016/j.jspi.2013.06.010.

Can efficiency be gained by correcting for misclassification?

Molin Wang^{a,c,b,1,*}, Xiaomei Liao^{b,c}, and Donna Spiegelman^{b,c}

^aChanning Division of Network Medicine, Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA

^bDepartment of Epidemiology, Harvard School of Public Health, Boston, MA, 02115, USA

^cDepartment of Biostatistics, Harvard School of Public Health, Boston, MA, 02115, USA

Abstract

This paper considers 2×2 tables arising from case-control studies in which the binary exposure may be misclassified. We found circumstances under which the inverse matrix method provides a more efficient odds ratio estimator than the naive estimator. We provide some intuition for the findings, and also provide a formula for obtaining the minimum size of a validation study such that the variance of the odds ratio estimator from the inverse matrix method is smaller than that of the naive estimator, thereby ensuring an advantage for the misclassification corrected result. As a corollary of this result, we show that correcting for misclassification does not necessarily lead to a widening of the confidence intervals, but, rather, in addition to producing a consistent estimate, can also produce one that is more efficient.

Keywords

2×2 table; Case control study; Misclassification; Odds ratio; Validation study design

1. Introduction

Exposure misclassification is a common problem in epidemiologic and other observational studies. Compared to scenarios where there is no misclassification of the exposure variable, exposure misclassification results in a loss of statistical power, often substantial (Lagakos, 1988; Tosteson and Tsiatis, 1988). Compared to the naive analysis, however, the picture is less clear. In methods widely used for bias correction of measures of association due to covariate measurement error and misclassification, such as the regression calibration method (Rosner, Spiegelman and Willett, 1990; Rosner, Willett and Spiegelman, 1989) and the matrix method (Greenland, 1988a), the variance of the corrected estimator can be partitioned into a sum of two components, in which one component is a function of the variance of the naive estimator divided by a de-attenuation factor typically less than 1, and the other a component related to the estimation of the parameters of the measurement error or misclassification model that has value always greater than 0 (Rosner, Spiegelman and Willett, 1990; Rosner, Willett and Spiegelman, 1989; Greenland, 1988a). Hence, these bias-

© 2013 Elsevier B.V. All rights reserved.

*Corresponding author. Phone: 1-617-4321412, Fax: 1-617-5667805.

¹The first author.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

correction procedures will typically yield estimates with variances greater than those to be obtained from the corresponding naive analysis, raising the question among practitioners about whether the cost of variance inflation due to measurement error or misclassification correction is worth the reduction in bias (Thiebaut, *et al.*, 2007). These concerns about variance inflation and loss of power have been a potential barrier to the application of the measurement error and misclassification bias correction methods which have been available for many years (Carroll *et al.*, 2006; Gustafson, 2003; Buonaccorsi, 2010). In this paper, we demonstrate that there is a broad range of scenarios likely to be encountered in practice where the variance of the odds ratio corrected for bias due to exposure misclassification is in fact smaller than the variance of the naive estimator. By doing so, we hope that one major barrier to more widespread application of this vast body of methodology into routine practice will be diminished.

The inverse matrix method (Marshall, 1990) is a well-known method for consistently estimating the odds ratio (OR) from a 2×2 table in a case-control study when the exposure is misclassified. When misclassification is differential, Lyles (2002) proved that the inverse matrix method is in fact the maximum likelihood method, but for non-differential misclassification, it may not be fully efficient. In this paper, we consider case-control studies where $Y (= 0, 1)$ denotes disease status, $X (= 0, 1)$ denotes the true exposure, $Z (= 0, 1)$ denotes the surrogate exposure, and information about the misclassification process is collected in a validation study, in which X, Z and Y are all available. Data in the main study are (Y_i, Z_i) , $i = 1, \dots, n_m$, and data in the validation study are (Y_i, Z_i, X_i) , $i = n_m + 1, \dots, n_m + n_v$ (Greenland, 1988b), where subscript i refers to subject i . Table 1 shows the layout of the 2×2 tables for the validation and main studies, respectively. In the table for the main study, the true cell counts a_m, b_m, c_m, d_m are unobserved. The corresponding cell counts in the full study are $A_f = A_m + 11, B_f = B_m + 10, C_f = C_m + 01, D_f = D_m + 00$,

$a_f = a_m + v'_{11}, b_f = b_m + v'_{10}, c_f = c_m + v'_{01}$ and $d_f = d_m + v'_{00}$. We consider two scenarios, one in which the validation study can be used to obtain information about the exposure-disease association, and one in which the validation study cannot be so used. The latter would occur when it is reasonable to assume that the misclassification parameters are transportable from the validation study to the main study but the OR is not transportable.

In this paper, we assume there is no misclassification of disease (i.e. Y is perfectly measured). The positive predictive value (PPV) and negative predictive value (NPV) are defined as $PPV_y = Pr(X = 1 | Z = 1, Y = y)$ and $NPV_y = Pr(X = 0 | Z = 0; Y = y)$, respectively.

We denote the inverse matrix estimator (Marshall, 1990) as $\widehat{OR}_I$. We have

$$\widehat{OR}_I = \hat{a}_h \hat{d}_h / (\hat{b}_h \hat{c}_h) \text{ with}$$

$$\begin{aligned} \hat{a}_h &= \widehat{PPV}_1 A_h + (1 - \widehat{NPV}_1) B_h, \hat{b}_h = (1 - \widehat{PPV}_1) A_h + \widehat{NPV}_1 B_h, \\ \hat{c}_h &= \widehat{PPV}_0 C_h + (1 - \widehat{NPV}_0) D_h, \hat{d}_h = (1 - \widehat{PPV}_0) C_h + \widehat{NPV}_0 D_h, \end{aligned}$$

where the subscript $h = f$ if the validation study is used to obtain information about the exposure-disease association, and $h = m$ otherwise. Note that to apply the inverse matrix method, PPV and NPV have to be estimated for cases and controls separately, whether misclassification is differential or non-differential. Therefore, the disease outcome, Y , must be available in the validation study, no matter whether the validation study is validly used to obtain information about the exposure-disease association or not.

In Section 2, we show that, when PPV and NPV are known, the variance of $\log(\widehat{OR}_I)$, based on the inverse matrix method, $\text{var}(\log(\widehat{OR}_I))$, is always smaller than or equal to the variance

of the naive estimator, $\text{var}(\log(\widehat{OR}_{naive}))$, where $\widehat{OR}_{naive} = A_h D_h / (B_h C_h)$. In addition, we provide some intuition for these findings. In Section 3, when PPV and NPV are not known and must be estimated from the validation study, we provide a formula for obtaining the minimum size of the validation study such that $\text{var}(\log(\widehat{OR}_I))$ is smaller than $\text{var}(\log(\widehat{OR}_{naive}))$, given fixed n_m or n_f . We illustrate our findings in a study of sudden infant death syndrome (SIDS) (Drews et al., 1990) in Section 4, and report on a numerical study investigating the significance of our findings under various scenarios in Section 5. We close with a brief discussion in Section 6.

2. The variance of the estimated odds ratio based on the inverse matrix method

The variance of $\log(\text{OR})$ estimator, $\text{var}(\log(\widehat{OR}_I))$, given by Marshall (1990), can be split into two parts, $\text{var}^{(1)}(\log(\widehat{OR}_I))$ and $\text{var}^{(2)}(\log(\widehat{OR}_I))$, where

$$\text{var}^{(1)}(\log(\widehat{OR}_I)) = \sum_{y=0}^1 \frac{(PPV_y + NPV_y - 1)^2 \text{var}[\widehat{Pr}(Z=1|Y=y)]}{[Pr(X=1|Y=y)(1 - Pr(X=1|Y=y))]^2} \quad (1)$$

is the variance for $\log(\widehat{OR}_I)$ assuming PPV_y and NPV_y are known, and $\text{var}^{(2)}(\log(\widehat{OR}_I))$ is the additional variance due to estimating PPV and NPV, which takes the form

$$\text{var}^{(2)}(\log(\widehat{OR}_I)) = \sum_{y=0}^1 [\delta_{1y} \text{var}(\widehat{PPV}_y) + \delta_{2y} \text{var}(\widehat{NPV}_y)]$$

with

$$\delta_{1y} = \frac{[Pr(Z=1|Y=y)]^2}{[Pr(X=1|Y=y)(1 - Pr(X=1|Y=y))]^2}, \quad \delta_{2y} = \frac{[1 - Pr(Z=1|Y=y)]^2}{[Pr(X=1|Y=y)(1 - Pr(X=1|Y=y))]^2},$$

$$\widehat{PPV}_y = v_{y11}/v_{y1}, \quad \widehat{NPV}_y = v_{y00}/v_{y0},$$

and $\text{var}[\widehat{Pr}(Z=1|Y=y)]$, $\text{var}(\widehat{PPV}_y)$ and $\text{var}(\widehat{NPV}_y)$ are calculated by the binomial variance formula. The above formula applies in both the situation where the validation study is used to obtain information about the exposure-disease association and those when it is not used (Marshall 1990; Lyles 2002). For presentational simplicity, throughout this section, let a, b, c, d denote a_h, b_h, c_h, d_h and A, B, C, D denote A_h, B_h, C_h, D_h , where $h = f$ if the validation study is used to obtain information about the exposure-disease association, and $h = m$ if it is not used. The ratio on the right hand side of (1) is the variance of $\log(\hat{a}/\hat{b})$ if $y = 1$ and $\log(\hat{d}/\hat{c})$ if $y = 0$.

Remark 1. For the inverse matrix method, when the PPV and NPV are known, $\text{var}(\log(\hat{a}/\hat{b}))$ is always less than or equal to $\text{var}(\log(A/B))$, and $\text{var}(\log(\hat{d}/\hat{c}))$ is always less than or equal to $\text{var}(\log(D/C))$. It then follows that $\text{var}^{(1)}(\log(\widehat{OR}_I)) \leq \text{var}(\log(AD/BC))$, where $\text{var}(\log(AD/BC))$ is the $\text{var}(\log(\widehat{OR}_{naive}))$, ignoring misclassification. The equality holds only if both $PPV_1 + NPV_1$ and $PPV_0 + NPV_0$ are 2 or 0. See the Appendix for a proof of Remark 1.

Next, we provide some insights for this counter-intuitive finding that correcting for bias due to misclassification can provide a more precise estimate than that obtained with the original misclassified naive estimate. When $PPV_1 + NPV_1 = 1$ and $PPV_0 + NPV_0 = 1$, from formula (1) we see that $\text{var}^{(1)}(\log(\widehat{OR}_I)) = 0$. Then $\hat{a} = PPV_1 A + (1 - NPV_1)B = PPV_1(A + B)$ and $b = NPV_1(A + B)$. Since for case-control studies, by design, the margin $A + B$ is fixed, both $\hat{a}$ and b above are fixed. That is, PPV_1 and NPV_1 contain all the information needed to obtain $\hat{a}$ and b when $PPV_1 + NPV_1 = 1$. Similarly, if $PPV_0 + NPV_0 = 1$, PPV_0 and NPV_0 contain all the information about d and $\hat{c}$. Intuitively, this then implies that if $PPV_1 + NPV_1$ is not equal to 1, they will contain some of the information about $\hat{a}$ and b . In fact, we can rewrite $\hat{a}$ as $\hat{a} = WA + (A + B)(1 - NPV_1)$, where the weight $W = PPV_1 + NPV_1 - 1$, and $-1 \leq W \leq 1$ ($W = 1$ implies no misclassification, and $W = -1$ implies every exposure value is misclassified). That is, $\hat{a}$ can be written as the sum of a weighted random part WA and a constant $(A + B)(1 - NPV_1)$. As long as the weight is not equal to 1 or -1 , it will reduce the variation in A . Similarly $b = WB + (A + B)(1 - PPV_1)$; that is, by knowing PPV_1 and NPV_1 , b can be split to a constant part and a downweighted random part.

3. Size of the validation study

Remark 1 implies that if the size of the validation study is large enough, $\text{var}(\log(\widehat{OR}_I))$ can be smaller than $\text{var}(\log(\widehat{OR}_{naive}))$. Next, we derive the minimum size of the validation study such that the variance of the estimate of $\log(OR)$, and thus the width of its confidence interval, is not increased by adjusting for bias due to misclassification. We focus on the situation where the validation study is used to obtain information about the exposure-disease association first, and then mention the result for the situation where the validation study is not so used at the end of this section.

By straightforward algebra,

$$\text{var}^{(1)}(\log(\widehat{OR}_I)) + \text{var}^{(2)}(\log(\widehat{OR}_I)) \leq \text{var}(\log(\widehat{OR}_{naive}))$$

is equivalent to

$$\frac{W_{01}}{v_{01}} + \frac{W_{00}}{v_{00}} + \frac{W_{11}}{v_{11}} + \frac{W_{10}}{v_{10}} \leq L \quad (2)$$

where $W_{y1} = PPV_y(1 - PPV_y)p_y^2/U_y$, $W_{y0} = NPV_y(1 - NPV_y)q_y^2/U_y$, with $p_y = P(Z = 1 | Y = y)$, $q_y = 1 - p_y$, $U_y = \{PPV_y p_y + (1 - NPV_y)q_y\}^2 \{NPV_y q_y + (1 - PPV_y)p_y\}^2$, $y = 0, 1$, $s_1 = A_f + B_f$, $s_0 = C_f + D_f$ and

$$L = \sum_{y=0}^1 \{1/(s_y p_y q_y) - (PPV_y + NPV_y - 1)^2 p_y q_y / (s_y U_y)\}. \text{ Our goal is to minimize the}$$

validation study sample size $n_v = \sum_{y,z=0}^1 v_{yz}$ with respect to v_{yz} , $y, z = 0, 1$, under restriction (2) and

$$v_{10} \leq m_{10}, v_{00} \leq m_{00}, v_{11} \leq m_{11}, v_{01} \leq m_{01}. \quad (3)$$

where $m_{11}, m_{10}, m_{01}, m_{00}$ are A_f, B_f, C_f and D_f respectively. Here, we have redefined the notation for those cell counts for convenience. This constrained minimization problem can be solved through the following steps, using the Lagrange method for minimization under inequality constraints.

Step 1. We ignore the constraints (3) and minimize n under constraint (2) only. We obtain

$$v_{yz}^* = W_{yz}^{1/2} \delta / L, y, z=0, 1, \text{ with } \delta = \sum_{y,z=0}^1 W_{yz}^{1/2}.$$

It follows that the minimum sample size of the validation study that makes the variance of the corrected logOR estimate no larger than the variance of uncorrected estimate is $n_v^* = \delta^2 / L$. Next, we check whether constraints (3) are satisfied for v_{yz}^* . If (3) holds, we have solved the constrained minimization problem; otherwise, it implies that one or more of the boundaries, m_{yz} in (3) have to be reached, and we need to pursue Steps 2 to 5 below.

Step 2. Assume that one boundary, say $m_{k_1 k_2}$ in (3) is reached; that is, $k_1 k_2 = m_{k_1 k_2}$ and ignore the other three inequalities in (3). By the Lagrange method, we obtain

$$v_{yz}^* = W_{yz}^{1/2} \delta_1, y, z=0, 1, yz \neq k_1 k_2 \text{ with } \delta_1 = \frac{\delta - W_{k_1 k_2}^{1/2}}{L - W_{k_1 k_2} / m_{k_1 k_2}}.$$

Similar to Step 1, we need to check whether the other three inequalities in constraints (3) are satisfied. If they are satisfied, we obtain the minimum sample size of the validation study

under constraints (2) and $k_1 k_2 = m_{k_1 k_2}$ as $n_{v_{k_1 k_2}}^* = \frac{(\delta - W_{k_1 k_2}^{1/2})^2}{L - W_{k_1 k_2} / m_{k_1 k_2}} + m_{k_1 k_2}$. If (3) is not satisfied, set $n_{v_{k_1 k_2}}^* = \infty$. Denote $n_v^{*(1)}$ as $\min(n_{v_{k_1 k_2}}^*, k_1, k_2=0, 1)$.

Step 3. Assume that two boundaries, say $m_{k_1 k_2}$ and $m_{j_1 j_2}$, in (3) are reached; that is, $k_1 k_2 = m_{k_1 k_2}$ and $j_1 j_2 = m_{j_1 j_2}$ and ignore the other two inequalities in (3). By the Lagrange method, we obtain

$$v_{yz}^* = W_{yz}^{1/2} \delta_1, y, z=0, 1, yz \neq k_1 k_2 \text{ or } j_1 j_2,$$

with

$$\delta_1 = \frac{\delta - W_{k_1 k_2}^{1/2} - W_{j_1 j_2}^{1/2}}{L - W_{k_1 k_2} / m_{k_1 k_2} - W_{j_1 j_2} / m_{j_1 j_2}}.$$

If the other two inequalities in constraints (3) are satisfied, we obtain the minimum sample size of the validation study under constraints (2), $k_1 k_2 = m_{k_1 k_2}$ and $j_1 j_2 = m_{j_1 j_2}$ as

$n_{v_{k_1 k_2, j_1 j_2}}^* = \frac{(\delta - W_{k_1 k_2}^{1/2} - W_{j_1 j_2}^{1/2})^2}{L - W_{k_1 k_2} / m_{k_1 k_2} - W_{j_1 j_2} / m_{j_1 j_2}} + m_{k_1 k_2} + m_{j_1 j_2}$. If (3) is not satisfied, set $n_{v_{k_1 k_2, j_1 j_2}}^* = \infty$. Denote $n_v^{*(2)}$ as $\min(n_{v_{k_1 k_2, j_1 j_2}}^*, k_1, k_2, j_1, j_2=0, 1, k_1 k_2 \neq j_1 j_2)$.

Step 4. Assume that only one boundary, say $m_{k_1 k_2}$, in (3) is not reached; that is, $yz = m_{yz}$ for $yz \neq k_1 k_2$, and ignore the inequality for $k_1 k_2$ in (3). By the Lagrange method, we obtain

$$v_{k_1 k_2}^* = \frac{W_{k_1 k_2}}{L - \sum_{y,z=0}^1 W_{yz} / m_{yz} + W_{k_1 k_2} / m_{k_1 k_2}}.$$

If $v_{k_1 k_2}^* \leq m_{k_1 k_2}$, we obtain $n_{v k_1 k_2}^{**} = v_{k_1 k_2}^* + \sum_{y \neq k_1 k_2} m_{yz}$; otherwise $n_{v k_1 k_2}^{**} = \infty$. Denote $n_v^{*(3)}$ as $\min(n_{v k_1 k_2}^{**}, k_1, k_2 = 0, 1)$.

Step 5. The minimum sample size of the validation study is $n_v^* = \min(n_v^{*(1)}, n_v^{*(2)}, n_v^{*(3)})$. If n_v^* is infinity, the variance of the corrected log(OR) estimate is always larger than that of the uncorrected estimate.

When the validation study is not used for estimating the exposure-disease association, the size of a validation study does not have to be less than or equal to that of the full study, eliminating constraints (3). Step 1 applies with the modification of replacing the full study margins s_y in the formula for L in (2) by the main study margins m_y , $y = 0, 1$, which is defined in Table 1.

4. An example

We illustrate these results in the case-control study investigating risk factors for sudden infant death syndrome (SIDS) (Drews et al., 1990). In this study, two controls were randomly selected from the birth certificates of infants born in six different populations. The first control was age-matched to the case, and the second control was matched by age, birth weight, and race. Mothers were interviewed regarding events which had occurred during their pregnancies, labors, and deliveries, and in the 5 weeks prior to the infant's death. To assess the extent of recall bias in the reporting of possible risk factors for SIDS, information about these same periods was abstracted from the medical records. Following Greenland (1988a,b) and Marshall (1990), here, we focus on antibiotic use as the exposure of interest.

There are 775 cases and 797 controls in the full study, with $A_f = 173$, $B_f = 602$, $C_f = 134$ and $D_f = 663$. There are 211 cases and 217 controls in the validation study. After excluding the observations in the validation study, the observed main study 2×2 table has 564 cases and 580 controls, with $A_m = 122$, $B_m = 442$, $C_m = 101$ and $D_m = 479$. From the validation study, PPV and NPV were estimated as 0.57 and 0.89, respectively, for cases, and 0.64 and 0.91, respectively, for controls.

Table 2 shows the results for uncorrected and corrected estimators. We see that, when using the inverse matrix method, the partial variance obtained under the assumption that the misclassification parameters are known, $\widehat{\text{var}}^{(1)}(\log(\widehat{OR}))$, is much smaller than the variance from the uncorrected estimate ignoring misclassification.

We applied the method in Section 3 to derive the minimum number of full study participants who need to be validated, such that the variance of the corrected log(OR) estimate is no larger than that of the uncorrected one. For Step 1, setting p_y , the prevalence of the exposure under possible misclassification, PPV, and NPV to their observed values in the original SIDS study, the minimum validation size is 1482. However, the constraints in (3) do not hold. By applying steps 2 to 5, we obtain the minimum validation study size 1556 with $n_{00} = 655$, $n_{01} = 134$, $n_{10} = 594$ and $n_{11} = 173$, where all the participants with $Z = 1$ in the full study are included in the validation study. This result is displayed in the last row in Table 2.

5. Numerical study

In the SIDS example above, almost all the participants in the primary study would need to be validated in order for the corrected estimator to have a variance less than or equal to that of the naive $\log(\widehat{OR})$. We conducted a numerical study to apply the method in Section 3 to a wide range of possible values for the odds ratio, PPVs and NPVs in scenarios likely to be

encountered in practice, to identify circumstances, if any, when the minimum required validation study size can be reasonably small. Please note that the asymptotic variances were directly calculated from the formulas; no simulations were conducted.

5.1. Numerical study design

We first focused on situations where the validation study was used to obtain information about the exposure-disease association. We assumed the numbers of controls and cases in the full study were 797 and 775, respectively, as in the SIDS study. Denote the exposure prevalence for X , $Pr(X=1|Y=y)$, as p_{yx} , $y=0, 1$. We specified p_{0x} , the exposure prevalence in the controls, representing the study base that produced the cases, and $OR_{Y|X}$, which represents the true OR for the X - Y relationship. Further, we assumed that the PPV_y and NPV_y , $y=0, 1$, were 0.9 or 0.8. Given PPV_y , NPV_y and p_{yx} , the exposure prevalence for Z , p_y , can be obtained by solving the equation $p_{yx} = PPV_y p_y + (1 - NPV_y)(1 - p_y)$, for $y=0, 1$, and $OR_{Y|Z}$, the OR for the Z - Y relationship, follows immediately from $OR_{Y|Z} = p_1(1 - p_0)/\{p_0(1 - p_1)\}$. Using the inverse matrix method estimator $\widehat{OR}_I$ to estimate $OR_{Y|X}$, the results given in Section 3 were applied to obtain the minimum percentage of full study participants who need to be validated in order for $\log(\widehat{OR}_I)$ to have a variance less than or equal to that of $\log(\widehat{OR}_{naive})$.

We also considered situations where the validation study was not used to obtain information about the exposure-disease association (please see the last paragraph of Section 3 for the algorithm for calculating the minimum validation study size for this scenario and how it differs from the algorithm for the scenario where the exposure-disease association information in the validation study is utilized). We assumed the numbers of controls and cases in the main study were 580 and 564, respectively, as in the SIDS study. We assumed the same prevalence p_{yx} of the correctly measured exposure, $OR_{Y|X}$, PPV_y and NPV_y , $y=0, 1$, as described above.

5.2. Numerical study results

Table 3 shows the results, giving the minimum percentage of full study participants who need to be validated in order for $\text{var}(\log(\widehat{OR}_I))$ to be less than or equal to that of $\log(\widehat{OR}_{naive})$. We calculated the minimum required validation study size for $OR_{Y|X}$ between 1.01 to 1.5 and p_{0x} equal to 0.2, 0.25 or 0.7, and found that when $OR_{Y|Z}$ was close to 1.0, the minimum required validation study size tended to be larger than that when $OR_{Y|Z}$ was very different from 1.0. For example, when $PPV_0 = NPV_1 = 0.8$, $PPV_1 = NPV_0 = 0.9$, and $OR_{Y|X}$ and p_{0x} were 1.01 and 0.2, respectively, validating more than only 2.2% of the full study led to a corrected OR estimate with a smaller variance than that of the naive estimate. Note that this case was very near the null, but with the PPVs and NPVs at 0.8 or 0.9, the naive analysis would lead to an extremely strong protective effect of $OR_{Y|Z} = 0.01$, underscoring the importance of correcting for misclassification even when the misclassification does not seem serious. In addition, as shown in Table 3, when $(OR_{Y|X}, p_{0x})$ equaled (1.5, 0.7) or (1.3, 0.25), the minimum required validation study sizes under $NPV_0 = PPV_0 = NPV_1 = PPV_1 = 0.8$ were smaller than that under $NPV_0 = PPV_0 = NPV_1 = PPV_1 = 0.9$. That is, to obtain an OR estimate with variance less than that of the uncorrected estimate, the counter-intuitive finding that the minimum required validation study size can decrease as misclassification increases is evident.

As shown in Table 3, we also calculated the proportion of the study sample in the validation study when the validation study was not used to obtain information about the exposure-

disease association. These ratios were even smaller than those in the situations where the validation study was used to obtain information about the exposure-disease association.

6. Discussion

We have demonstrated that when the validation study is large enough, $\text{var}(\log(\widehat{OR}_I))$ can be smaller than $\text{var}(\log(\widehat{OR}_{naive}))$. Wald confidence intervals around the point estimate will thus be narrower than those around the naive estimate. This result dispels a common misperception that the confidence interval is invariably wider when misclassification is adjusted for, compared to the naive analysis. This result applies to both the differential and nondifferential misclassification cases, and holds even when the validation study is not used to obtain information about the exposure-disease association. However, if estimation and inference are based on the matrix method (Barron, 1977; Greenland, 1988a), these results do not apply. Rather, the variance of the $\log(OR)$ estimate obtained by the matrix method is always greater than or equal to $\text{var}(\log(\widehat{OR}_{naive}))$ (proof available from first author upon request for the differential misclassification case; for non-differential misclassification this result has appeared previously (Greenland and Gustafson, 2006)), even when the sensitivity and specificity are known. In addition, as shown in Section 5, we also found that

$\text{var}(\log(\widehat{OR}_I))$ does not always increase as misclassification increases. This finding is consistent with that for the Berkson measurement error model studied by Demidenko and Spiegelman (1997), where the authors showed that, for a continuous exposure, one can decrease the variance of the maximum likelihood estimate by increasing measurement error.

We noted that in Table 2, the percent of study subjects in the validation study needed in order to obtain a variance equivalent to the uncorrected was 99%. In the particular case of the SIDS study (Drews et al., 1990), it is unlikely that such a design could be contemplated, given the likely higher cost of the gold standard measurements relative to those of the surrogate. However, it was found in Table 3, that under a wide range of other scenarios likely to be encountered in practice, the percent of study subjects in the validation study can be rather small and well within the range of what is realistic, in order to obtain an equivalent efficiency, or better, relative to that of the naive analysis.

When the exposure is validated in a sub-sample of the primary study, Greenland (1988b) developed methods for determining the optimal validation proportion by minimizing the variance of $\log(OR)$ estimate subject to a fixed cost constraint based upon the variance of the matrix method estimator. This optimal validation proportion depends upon the ratio of the cost per subject of obtaining the error-prone measurements over that per subject of obtaining the validation measurement. Our findings can also be used to provide insight on optimal study design. For the inverse matrix method, the variance due to estimating the misclassification parameters, $\text{var}^{(2)}(\log(\widehat{OR}_I))$, is often a major contributor to the total variance of $\log(\widehat{OR}_I)$. In contrast, for the matrix method, both the component of the variance due to estimating the misclassification parameters and the component due to estimating odds ratio assuming the misclassification parameters are known are often major contributors to the total variance of $\log(OR)$ estimate. Therefore, if using the inverse matrix method, it may be more beneficial for the validation study to be large since $\text{var}(\widehat{PPV}_y)$ and $\text{var}(\widehat{NPV}_y)$ will be minimized, and thus the total variance. As shown in Table 3, even with a small validation study, the inverse matrix corrected estimator can achieve good precision.

The power for $H_0 : OR_{Y|X} = 1$ versus the alternative $H_A : OR_{Y|X} \neq 1$ to detect an odds ratio of $OR_{Y|X} = r$ or greater may be calculated by

$\prod_{naive}(r^*) = 1 - \Phi\left(z_{1-\alpha/2} - |\log(r^*)|/se\{\log(\widehat{OR}_{naive})\}\right)_{OR_{Y|Z}=r^*}$ for a given Type I error rate when the Wald test statistic is based upon the naive analysis, where $\Phi(\cdot)$ is the cumulative standard normal distribution function and

$r^* = \frac{NPV_1(1 - p_{0x} + rp_{0x}) + p_{0x} - 1}{PPV_1(1 - p_{0x} + rp_{0x}) - rp_{0x}} \cdot \frac{PPV_0 - p_{0x}}{NPV_0 + p_{0x} - 1}$. On the other hand, the power for $H_0: OR_{Y|X} = 1$ versus the alternative $H_A: OR_{Y|X} \neq 1$ may be calculated by

$\prod_I(r) = 1 - \Phi\left(z_{1-\alpha/2} - |\log(r)|/se\{\log(\widehat{OR}_I)\}\right)_{OR_{Y|X}=r}$ when the inverse matrix method is used as the basis of the Wald test statistic. Under nondifferential misclassification with a binary exposure, since r is always further away from the null than r^* , $\prod_I(r)$ will be greater than $\prod_{naive}(r^*)$ as long as $\text{var}(\log(\widehat{OR}_I)) \leq \text{var}(\log(\widehat{OR}_{naive}))$, as can be obtained by design using the algorithm in Section 3. It should be noted that under differential misclassification, hypothesis testing based on the naive method may not be valid, since the type I error rate may not equal to

This paper focused on misclassification in the context of case-control studies. In cohort studies, the exposure is usually validated without a specific disease outcome in mind which in any case has not yet occurred, and when the disease of interest is identified, it is often not possible to estimate the PPV and NPV in the cases due to small numbers in the validation study. Therefore, the inverse matrix method can rarely be used for cohort studies.

Most importantly, we have shown that correcting for misclassification does not necessarily lead to wider confidence intervals compared to the naive analysis, but, rather, in addition to producing a consistent estimate, may also produce one that is more efficient.

Acknowledgments

We acknowledge the funding support of NIH R01 CA 50597 in performing this work. We are also thank the Associate Editor and two referees for their useful comments that improved the manuscript.

Appendix

Proof of Remark 1. When ignoring misclassification (i.e. $X = Z$, $PPV_1 = NPV_1 = 1$, $PPV_0 = NPV_0 = 1$), $\text{var}(\log(\widehat{OR}_I))$ reduces to

$$\frac{\text{var}[\widehat{Pr}(Z=1|Y=1)]}{[Pr(Z=1|Y=1)(1 - Pr(Z=1|Y=1))]^2} + \frac{\text{var}[\widehat{Pr}(Z=1|Y=0)]}{[Pr(Z=1|Y=0)(1 - Pr(Z=1|Y=0))]^2} = \text{var}(\log(A/B)) + \text{var}(\log(D/C)),$$

which is $\text{var}(\log(\widehat{OR}_{naive}))$. The following proof is for $y = 1$, which applies similarly for $y = 0$. Denote $Pr(Z = 1|Y = y)$ as p_y and $Pr(Z = 0|Y = y)$ as q_y . First we show that

$$\frac{(PPV_1 + NPV_1 - 1)^2 \text{var}[\widehat{Pr}(Z=1|Y=1)]}{[Pr(X=1|Y=1)(1 - Pr(X=1|Y=1))]^2} \leq \frac{\text{var}[\widehat{Pr}(Z=1|Y=1)]}{[p_1 q_1]^2}, \quad (\text{A.1})$$

which is equivalent to

$$\frac{|PPV_1 + NPV_1 - 1|}{\{PPV_1 p_1 + (1 - NPV_1) q_1\} \{(1 - PPV_1) p_1 + NPV_1 q_1\}} \leq \frac{1}{p_1 q_1}.$$

The inequality above can be rewritten as

$$\{PPV_1 p_1 + (1 - NPV_1) q_1\} \{ (1 - PPV_1) p_1 + NPV_1 q_1 \} - |PPV_1 + NPV_1 - 1|_{p_1 q_1} \geq 0 \quad (A.2)$$

which holds if $PPV_1 + NPV_1 - 1 = 0$. If $PPV_1 + NPV_1 < 1$, the l.h.s. of (A.2) can be rewritten as

$$PPV_1(1 - PPV_1)p_1^2 + NPV_1(1 - NPV_1)q_1^2 + 2PPV_1NPV_1p_1q_1,$$

which is non-negative since all three terms are non-negative. If $PPV_1 + NPV_1 > 1$, the expression above can be written as

$$\begin{aligned} & PPV_1(1 - PPV_1)p_1^2 \\ & + NPV_1(1 - NPV_1)q_1^2 \\ & + 2(1 - PPV_1)(1 - NPV_1)p_1q_1 > (1 - PPV_1)(1 - NPV_1)(p_1^2 + q_1^2) + 2(1 - PPV_1)(1 - NPV_1)p_1q_1 \\ & = (1 - PPV_1)(1 - NPV_1)(p_1 + q_1)^2, \end{aligned}$$

which is, again, non-negative. Therefore, inequality (A.2) holds, and it follows that inequality (A.1) holds.

References

1. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977; 33:414–418. [PubMed: 884199]
2. Buonaccorsi, JP. Measurement error: models, methods and applications. 2010. Edited by Boca Raton, FL, Chapman and Hall.
3. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Crainiceanu CM: Measurement Error in Nonlinear Models. 2006. Edited by London, Chapman & Hall.
4. Demidenko E, Spiegelman D. A paradox: more measurement error can lead to more efficient estimates. *Commun. Statist. - Theory Meth*. 1997; 26:1649–1675.
5. Drews CD, Kraus JF, Greenland S. Recall bias in a case-control study of sudden-infant-death-syndrome. *International J. Epidemiol*. 1990; 19:405–411.
6. Greenland S. Variance-estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*. 1988a; 7:745–757. [PubMed: 3043623]
7. Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. *Journal of Clinical Epidemiology*. 1988b; 41:1167–1174. [PubMed: 3062138]
8. Greenland S, Gustafson G. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *American Journal of Epidemiology*. 2006; 164:64–68.

9. Gustafson, P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. 2003. Edited by Chapman and Hall CRC.
10. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*. 1988; 7:257–274. [PubMed: 3353607]
11. Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*. 2002; 58:1034–1037. [PubMed: 12495160]
12. Marshall R. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J. Clin. Epidemiol.* 1990; 43:941–947. [PubMed: 2213082]
13. 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; 132:734–745. [PubMed: 2403114]
14. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med.* 1989; 8:1051–1069. discussion 1071-1053. [PubMed: 2799131]
15. Thiebaut ACM, Freedman LS, Carroll RJ, Kipnis V. Is it necessary to correct for measurement error in nutritional epidemiology? *Annals of Internal Medicine*. 2007; 146:65-W10. [PubMed: 17200225]
16. Tosteson TD, Tsiatis AA. The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*. 1988; 75:507–517.

Table 1

Data layout for a case-control study with misclassified exposure data

		Validation study				Main study				
		Y = 1		Y = 0		Observed		Not observed		
Z	X	X = 1	X = 0	X = 1	X = 0	Y	Z = 1	Z = 0	X = 1	X = 0
1	111	110	11	011	010	01	A_m	B_m	a_m	b_m
0	101	100	10	001	000	00	C_m	D_m	c_m	d_m
	v_{11}	v_{10}		v_{01}	v_{00}					n_m

Analysis results of the SIDS data (Drews et al., 1990)

Table 2

Design	Method	MS size	VS size	$\widehat{OR}$	$\widehat{\text{var}}(\log(\widehat{OR}))$	$\widehat{\text{var}}^{(1)}(\log(\widehat{OR}))$	$\widehat{\text{var}}^{(2)}(\log(\widehat{OR}))$
MS	Uncorrected	1144	0	1.31	0.024	NA	NA
Full	Uncorrected	1572	0	1.42	0.016	NA	NA
IVS	Inverse matrix	1144	428	1.21	0.049	0.004	0.045
EVS	Inverse matrix	1144	428	1.16	0.051	0.004	0.047
Optimal MS/IVS [†]	Inverse matrix	16	1556	1.21	0.016	0.004	0.012

$\widehat{PPV}_1=0.57, \widehat{NPV}_1=0.89, \widehat{PPV}_0=0.64, \widehat{NPV}_0=0.91, p1=0.22, p0=0.17.$

MS refers to main study; IVS and EVS refer to the situations where the validation study is used and not used to obtain information about the exposure-disease relationship, respectively.

[†]Optimal MS/IVS is the design with minimum validation study size such that the variance of the corrected OR estimate is no larger than that corresponding to the uncorrected one.

Table 3

Minimum percentage of participants in the validation study such that $\text{var}(\log(\widehat{OR}_I)) \leq \text{var}(\log(\widehat{OR}_{naive}))$.

For X		For Z^{**}				$n/(n + n_m)\%$		
$OR_{Y X}$	p_{0c}^*	OR_{YZ}^*	p_0^*	PPV_0	NPV_0	PPV_1	NPV_1	$EVS^\dagger$
1.5	0.7	5.20	0.83	0.8	0.8	0.8	0.8	26.2
		1.11	0.83	0.8	0.8	0.9	0.9	60.2
		8.67	0.75	0.9	0.9	0.8	0.8	22.9
		5.08	0.86	0.8	0.9	0.8	0.9	21.8
1.3	0.25	1.89	0.71	0.9	0.8	0.9	0.8	79.9
		0.79	0.86	0.8	0.9	0.9	0.8	59.4
		12.2	0.71	0.9	0.8	0.8	0.9	21.2
		1.85	0.75	0.9	0.9	0.9	0.9	65.9
1.1	0.2	2.26	0.08	0.8	0.8	0.8	0.8	44.7
		3.72	0.08	0.8	0.8	0.9	0.9	42.8
		0.89	0.19	0.9	0.9	0.8	0.8	62.7
		1.49	0.21	0.8	0.9	0.8	0.9	86.7
1.01	0.2	2.23	0.07	0.9	0.8	0.9	0.8	37.2
		0.63	0.21	0.8	0.9	0.9	0.8	61.0
		5.29	0.07	0.9	0.8	0.8	0.9	40.2
		1.47	0.19	0.9	0.9	0.9	0.9	71.9
1.01	0.2	0.19	0.13	0.9	0.9	0.8	0.8	19.6
		1.19	0.14	0.8	0.9	0.8	0.9	68.8
		0.14	0.14	0.8	0.9	0.9	0.8	18.0
		1.18	0.13	0.9	0.9	0.9	0.9	56.8
1.01	0.2	0.02	0.13	0.9	0.9	0.8	0.8	2.5
		1.02	0.14	0.8	0.9	0.8	0.9	66.5
		0.01	0.14	0.8	0.9	0.9	0.8	2.2
		1.02	0.13	0.9	0.9	0.9	0.9	55.1

* $p_{0X} = P(X=1|Y=0)$, $p_0 = P(Z=1|Y=0)$.

^{***} $OR_{Y|Z}$ and ρ_0 are fixed once $OR_{Y|X}$, ρ_{0X} , PPV 's and NPV 's are set.

^{***} IVS refers to the situations where the validation study is used to obtain information about the exposure-disease association; the full study size is $n + n_M = 1572$ with 797 cases and 775 controls.

[†] EVS refers to situations where the study is not used to obtain information about the exposure-disease association; the main study size $n_M = 1144$ with 564 cases and 580 controls.

^{††} The PPV 's and NPV 's combinations that are inconsistent with given $OR_{Y|X}$ and ρ_{0X} are excluded.