Journal of Data Science **9**(2011), 529-548

Confidence Intervals for the Risk Ratio Using Double Sampling with Misclassified Binomial Data

Dewi Rahardja[1*] and Dean M. Young[2]
[1] *University of Texas Southwestern Medical Center and* [2] *Baylor University*

Abstract: We derive three likelihood-based confidence intervals for the risk ratio of two proportion parameters using a double sampling scheme for misclassified binomial data. The risk ratio is also known as the relative risk. We obtain closed-form maximum likelihood estimators of the model parameters by maximizing the full-likelihood function. Moreover, we develop three confidence intervals: a naive Wald interval, a modified Wald interval, and a Fieller-type interval. We apply the three confidence intervals to cervical cancer data. Finally, we perform two Monte Carlo simulation studies to assess and compare the coverage probabilities and average lengths of the three interval estimators. Unlike the other two interval estimators, the modified Wald interval always produces close-to-nominal confidence intervals for the various simulation scenarios examined here. Hence, the modified Wald confidence interval is preferred in practice.

Key words: Binomial data, double sampling, misclassification, relative risk, risk ratio.

1. Introduction

Bross (1954) was the first among several researchers who have studied the effect of misclassification on the classical proportion estimators. In general, two types of misclassification for binary misclassified observations exist: false-positive and false-negative binary observations. For example, visual inspection by a midwife or obstetrician may erroneously classify a normal child as having Down's syndrome (false-positive), or one may classify a child with Down's syndrome as being healthy (false-negative). In many applications with misclassified binary data, both misclassification types are present.

Because classical estimators that ignore misclassification are biased, one needs additional data to correct the bias and achieve model identifiability. Various methods in the statistical literature have been proposed for this purpose. For the

*Corresponding author.

Bayesian paradigm, when an infallible classifier is unavailable or prohibitively expensive, one can use sufficiently informative priors to obtain model identifiability. Another information-producing method is to use multiple fallible classifiers. This article focuses on an information-addition method first proposed by Tenenbein (1970) that includes training data obtained by double sampling.

One can apply Tenenbein's double sampling scheme when both fallible and infallible measuring devices or classifiers are available. Usually, the fallible classifier is relatively inexpensive but may misclassify units, while the infallible classifier is generally much more expensive but is infallible. Tenenbein's approach was to compromise between the two extremes by using the infallible classifier on only a small portion of the data and using the fallible classifier on all of the data. This approach, called double sampling, not only enables model identifiability but is also economical.

A number of researchers have used misclassified binary data to provide point and interval estimation methods for various functions of the proportion parameters of interest. For one-sample binomial problems where only one type of error or misclassification is present, Lie, Heuch and Irgens (1994) have used a maximum likelihood approach, where false-negative errors are corrected with multiple fallible classifiers, whereas York, Madigan, Heuch and Lie (1995) have considered the same problem from a Bayesian approach. Using data obtained by double sampling, Moors, van der Genugten and Strijbosch (2000) have discussed method of moments and maximum likelihood estimation, in addition to one-sided interval estimation. Also, Boese, Young and Stamey (2006) have derived several likelihood-based confidence intervals (CIs) for a single proportion parameter, while Lee and Byun (2008) have provided Bayesian credible intervals using noninformative priors for the same problem.

Additionally, several researchers have studied one-sample problems with both types of binary misclassification errors. In conjunction with double sampling, Tenenbein (1970) has proposed a maximum likelihood estimator for a single proportion parameter and has derived an expression for the estimator's asymptotic variance. For the case when training data are unavailable in the one-sample problem, Gaba and Winkler (1992) and Viana, Ramakrishnan and Levy (1993) have developed Bayesian approaches using sufficiently informative priors.

For the two-sample problem with both types of binary misclassification errors, Bayesian inference methods using sufficiently informative priors have also been developed when training data are unavailable. For example, see Evans, Guttman, Haitovsky and Swartz (1996) for risk-difference estimation, that is, the difference of two proportion parameters, and Gustafson, Le and Saskin (2001) for estimation of odds ratios. When training data is obtained through double sampling, Boese (2003) has derived several likelihood-based CIs for the risk difference.

So far, no inference methods for the risk ratio of two proportion parameters have been published for two-sample misclassified binary data. Such inference methods have applications in many fields including epidemiology and marketing. For example, Hildesheim, Mann, Brinton, Szklo, Reeves and Rawls (1991) reported a study assessing the relationship between exposure to herpes simplex virus (HSV) and invasive cervical cancer (ICC). The western blot procedure was treated as a fallible detector of HSV and was applied to every woman in the study to detect the exposure to HSV. A sub-sample of the women were also tested using the refined western blot procedure, which is a relatively accurate procedure and, thus, was treated as infallible. The estimation of the risk ratio is of interest for this data to explore the association between exposure to HSV and having ICC.

In this article, we develop point and interval estimators for this problem. The remainder of this paper is organized as follows. In Section 2 we describe the data, and in Section 3 we derive three likelihood-based interval estimators of a risk ratio using double sampling with misclassified data containing both false-negative and false-positive observations. In Section 4 we illustrate the newly derived interval estimators using real cervical cancer data. We examine and compare the performance of three interval estimators in Section 5, and we give a brief discussion in Section 6.

2. The Data

In this section we introduce notation and rigorously describe two-sample misclassified binomial data. The original data are obtained with a fallible classifier that produces both false-positive and false-negative observations.

We first introduce notation necessary for describing the data. Let F_{ij} be the observed classification by the fallible classifier for the jth observation unit in the ith sample, where $i = 1, 2$, $j = 1, \cdots, M_i$, and

$$F_{ij} = \begin{cases} 1, & \text{if the result by the fallible classifier is positive,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $X_i = \sum_j F_{ij}$ and $Y_i = M_i - X_i$ be the observed number of positive and negative observations, respectively. The data obtained by the fallible classifier for sample i, $i = 1, 2$, are displayed in Table 1.

Table 1: Data from the Fallible Classifier for Sample i, $i = 1, 2$

Classification	0	1	Total
Count	Y_i	X_i	M_i

Similarly, we define the true classification of the jth observation unit in the ith sample as

$$T_{ij} = \begin{cases} 1, & \text{if the classifier result is truly positive,} \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, the T_{ij} may not be observable and misclassification occurs when $T_{ij} \neq F_{ij}$.

Also, we let

$$\begin{aligned} p_i &\equiv \Pr(T_{ij} = 1), \\ \pi_i &\equiv \Pr(F_{ij} = 1), \\ \phi_i &\equiv \Pr(F_{ij} = 1 | T_{ij} = 0), \end{aligned}$$

and

$$\theta_i \equiv \Pr(F_{ij} = 0 | T_{ij} = 1).$$

Here, p_i is the actual proportion parameter of interest, π_i is the proportion parameter of the fallible classifier, ϕ_i and θ_i are the false-positive and the false-negative rates, respectively, for the fallible classifier. Note that we allow the false-positive rates and false-negative rates to be different between the two samples, i.e., we allow $\phi_1 \neq \phi_2$ and $\theta_1 \neq \theta_2$. Also we remark that π_1 and π_2 are not additional unique parameters because

$$\begin{aligned} \pi_i &= \Pr(T_{ij} = 1) \Pr(F_{ij} = 1 | T_{ij} = 1) + \Pr(T_{ij} = 0) \Pr(F_{ij} = 1 | T_{ij} = 0) \\ &= p_i(1 - \theta_i) + q_i \phi_i, \end{aligned} \tag{2.1}$$

where $q_i = 1 - p_i$. As noted in Section 1, we wish to develop point and interval estimators of the risk ratio

$$r = p_1 / p_2. \tag{2.2}$$

Because π_i is determined through p_i, ϕ_i, and θ_i, $i = 1, 2$, effectively six parameters result in the model: p_1, ϕ_1, θ_1, p_2, ϕ_2, θ_2. However, the sufficient statistics dimension is only two because X_1 and X_2 are the minimal sufficient statistics for this model. Therefore, six parameters in model (2.1) are unidentifiable because the dimension of the sufficient statistics is less than the number of parameters and, therefore, additional data are needed for model identifiability. In this paper we use double sampling to provide additional information. Specifically, in addition to the original fallible data classified only by the fallible classifier, new but smaller training data are obtained when classifying each observation unit in this training data by both the fallible and the infallible classifiers. The double sampling paradigm has attracted researchers' interests due to its practicality.

For example, Tenenbein (1972) considered parameter estimation and sample size calculation in quality control problems. Hochberg (1977) studied a model for misclassified binomial data where covariates were adjusted for inference. Boese *et al.* (2006) reported several likelihood-based methods for constructing confidence intervals for a one-sample proportion.

In this paper we assume that for the ith sample, training data of size n_i are obtained using double sampling in addition to the original fallible data of size M_i, $i = 1, 2$. Hence, the size of the combined data is $N_i = M_i + n_i$ for sample i. Table 2 presents the combined data by concatenating the original and training data. In Table 2 we use n_{ijk} to denote the number of observation units classified as j and k by the infallible and fallible classifiers, respectively. For example, n_{i01} is the number of observation units in the ith sample classified as negative by the infallible classifier but classified as positive by the fallible classifier. With the additional training data, the dimension of the sufficient statistic for the combined data is sufficient for estimating all parameters and, therefore, the full model is identifiable. For future estimation methodology development, we present the cell probabilities corresponding to Table 2 in Table 3.

Table 2: Data for Sample i

		Fallible Classifier		
Data	Infallible Classifier	0	1	Total
Training	0	n_{i00}	n_{i01}	$n_{i0\cdot}$
	1	n_{i10}	n_{i11}	$n_{i1\cdot}$
	Total	$n_{i\cdot 0}$	$n_{i\cdot 1}$	n_i
Original	NA	Y_i	X_i	M_i

NA: Not Available

Table 3: Cell Probabilities for Sample i

		Fallible Classifier		
Data	Infallible Classifier	0	1	Total
Training	0	$q_i(1-\phi_i)$	$q_i\phi_i$	q_i
	1	$p_i\theta_i$	$p_i(1-\theta_i)$	p_i
Original	NA	$1-\pi_i$	π_i	1

NA: Not Available

3. The Model

For data described in the previous section, we derive point and interval estimators for the risk ratio (2.2) of two proportion parameters using double sampling on possibly misclassified data. In particular, we derive closed-form maximum likelihood estimators (MLEs). In addition, we obtain an asymptotic covariance matrix of the vector of MLEs by computing the inverse of the Fisher information matrix. Finally, we develop two closed-form Wald-based CIs and a Fieller-type CI for the risk ratio r based on the full likelihood.

3.1 The Full Likelihood Function

Table 2 presents the data for the inference problem under consideration. For sample i, the observed counts $(n_{i00}, n_{i01}, n_{i10}, n_{i11})'$ of the training data have a quadrinomial distribution with total size n_i and probabilities displayed in an upper right 2×2 submatrix in Table 3, i.e.,

$$(n_{i00}, n_{i01}, n_{i10}, n_{i11})|p_i, \phi_i, \theta_i \sim \text{Quad}[n_i, (q_i(1-\phi_i), q_i\phi_i, p_i\theta_i, p_i(1-\theta_i))], \quad (3.1)$$

where Quad is an abbreviation for the Quadrinomial distribution. In addition, the observed counts (X_i, Y_i) have the binomial distribution

$$(X_i, Y_i)|p_i, \phi_i, \theta_i \sim \text{Bin}[M_i, (\pi_i, 1-\pi_i)]. \quad (3.2)$$

Because $(n_{i00}, n_{i01}, n_{i10}, n_{i11})'$ and $(X_i, Y_i)'$ are independent for sample i and because sample 1 is independent of sample 2, the probability density function of the data vector given the parameter vector is

$$f(\boldsymbol{d}|\boldsymbol{\eta}) \propto \prod_{i=1}^{2}\{[q_i(1-\phi_i)]^{n_{i00}}(q_i\phi_i)^{n_{i01}}(p_i\theta_i)^{n_{i10}}[p_i(1-\theta_i)]^{n_{i11}}\pi_i^{X_i}(1-\pi_i)^{Y_i}\}, \quad (3.3)$$

where

$$\boldsymbol{d} = (n_{100}, n_{101}, n_{110}, n_{111}, X_1, Y_1, n_{200}, n_{201}, n_{210}, n_{211}, X_2, Y_2)' \quad (3.4)$$

and

$$\boldsymbol{\eta} = (p_1, \phi_1, \theta_1, p_2, \phi_2, \theta_2)'.$$

Finally, we can express the full likelihood function as

$$L_f(\boldsymbol{\eta}) \propto \prod_{i=1}^{2}\{[q_i(1-\phi_i)]^{n_{i00}}(q_i\phi_i)^{n_{i01}}(p_i\theta_i)^{n_{i10}}[p_i(1-\theta_i)]^{n_{i11}}\pi_i^{X_i}(1-\pi_i)^{Y_i}\}. \quad (3.5)$$

3.2 MLEs Based on the Full Likelihood Function

We now derive the maximum likelihood estimators (MLEs) of all parameters of interest. Generally, directly maximizing (3.5) with respect to $\boldsymbol{\eta}$ requires such numerical methods as the Newton-Raphson algorithm. These numerical methods are computationally expensive and may have convergence issues. Instead of using these numerical methods, we first perform a reparameterization of parameters $\boldsymbol{\eta}$ and then derive closed-form solutions. Let

$$\lambda_{i1} \equiv p_i(1-\theta_i)/\pi_i, \tag{3.6}$$

$$\lambda_{i2} \equiv p_i\theta_i/(1-\pi_i), \tag{3.7}$$

and $\boldsymbol{\gamma} \equiv (\lambda_{11}, \lambda_{12}, \pi_1, \lambda_{21}, \lambda_{22}, \pi_2)'$, $i = 1, 2$. Using (2.1), (3.6), and (3.7), we see that (3.5) can be reexpressed as

$$L_f(\boldsymbol{\gamma}) \propto \prod_{i=1}^{2} \left[\lambda_{i1}^{n_{i11}}(1-\lambda_{i1})^{n_{i01}}\lambda_{i2}^{n_{i10}}(1-\lambda_{i2})^{n_{i00}}\pi_i^{X_i+n_{i\cdot 1}}(1-\pi_i)^{Y_i+n_{i\cdot 0}}\right]. \tag{3.8}$$

Therefore, the full log likelihood is

$$\begin{aligned} l_f(\boldsymbol{\gamma}) \propto \sum_{i=1}^{2} [&n_{i11}\log\lambda_{i1} + n_{i01}\log(1-\lambda_{i1}) + n_{i10}\log\lambda_{i2} + n_{i00}\log(1-\lambda_{i2}) + \\ &(X_i + n_{i\cdot 1})\log\pi_i + (Y_i + n_{i\cdot 0})\log(1-\pi_i)], \end{aligned} \tag{3.9}$$

and the corresponding score vector is

$$\begin{aligned} \boldsymbol{s}_f(\boldsymbol{\gamma}) &\equiv \frac{\partial l_f(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \\ &= \left[\frac{n_{111}}{\lambda_{11}} - \frac{n_{101}}{1-\lambda_{11}}, \frac{n_{110}}{\lambda_{12}} - \frac{n_{100}}{1-\lambda_{12}}, \frac{X_1 + n_{1\cdot 1}}{\pi_1} - \frac{Y_1 + n_{1\cdot 0}}{1-\pi_1},\right. \\ &\quad \left.\frac{n_{211}}{\lambda_{21}} - \frac{n_{201}}{1-\lambda_{21}}, \frac{n_{210}}{\lambda_{22}} - \frac{n_{200}}{1-\lambda_{22}}, \frac{X_2 + n_{2\cdot 1}}{\pi_2} - \frac{Y_2 + n_{2\cdot 0}}{1-\pi_2}\right]'. \end{aligned} \tag{3.10}$$

We obtain the MLE for $\boldsymbol{\gamma}$ by setting $\boldsymbol{s}_f(\boldsymbol{\gamma}) = \mathbf{0}$ and solving for λ_{i1}, λ_{i2}, and π_i, so that

$$\hat{\lambda}_{i1} = \frac{n_{i11}}{n_{i\cdot 1}},$$

$$\hat{\lambda}_{i2} = \frac{n_{i10}}{n_{i\cdot 0}},$$

and

$$\hat{\pi}_i = \frac{X_i + n_{i\cdot 1}}{N_i},$$

$i = 1, 2$. By solving (2.1), (3.6), and (3.7) and applying the invariance property of MLEs, we find the MLEs for $\boldsymbol{\eta}$ are

$$\hat{p}_i = \hat{\pi}_i \hat{\lambda}_{i1} + (1 - \hat{\pi}_i)\hat{\lambda}_{i2},$$

$$\hat{\phi}_i = (1 - \hat{\lambda}_{i1})\hat{\pi}_i / \hat{q}_i,$$

$$\hat{\theta}_i = \hat{\lambda}_{i2}(1 - \hat{\pi}_i)/\hat{p}_i,$$

$i = 1, 2$, and

$$\hat{r} = \hat{p}_1 / \hat{p}_2. \tag{3.11}$$

3.3 The Full Likelihood Information Matrix

From (3.10), the Hessian matrix is

$$\begin{aligned} \boldsymbol{H}_f(\boldsymbol{\gamma}) = \text{Diag} \Bigg[& -\frac{n_{111}}{\lambda_{11}^2} - \frac{n_{101}}{(1-\lambda_{11})^2}, -\frac{n_{110}}{\lambda_{12}^2} - \frac{n_{100}}{(1-\lambda_{12})^2}, \\ & -\frac{X_1 + n_{1\cdot 1}}{\pi_1^2} - \frac{Y_1 + n_{1\cdot 0}}{(1-\pi_1)^2}, -\frac{n_{211}}{\lambda_{21}^2} - \frac{n_{201}}{(1-\lambda_{21})^2}, \\ & -\frac{n_{210}}{\lambda_{22}^2} - \frac{n_{200}}{(1-\lambda_{22})^2}, -\frac{X_2 + n_{2\cdot 1}}{\pi_2^2} - \frac{Y_2 + n_{2\cdot 0}}{(1-\pi_2)^2} \Bigg]. \end{aligned} \tag{3.12}$$

Thus, the expected Fisher information matrix is

$$\begin{aligned} \boldsymbol{I}_f(\boldsymbol{\gamma}) = \text{Diag} \Bigg[& \frac{n_1 \pi_1}{\lambda_{11}(1-\lambda_{11})}, \frac{n_1(1-\pi_1)}{\lambda_{12}(1-\lambda_{12})}, \frac{N_1}{\pi_1(1-\pi_1)}, \\ & \frac{n_2 \pi_2}{\lambda_{21}(1-\lambda_{21})}, \frac{n_2(1-\pi_2)}{\lambda_{22}(1-\lambda_{22})}, \frac{N_2}{\pi_2(1-\pi_2)} \Bigg]. \end{aligned}$$

Because the necessary regularity conditions are satisfied for this model, the MLE vector $\hat{\boldsymbol{\gamma}} = (\hat{\lambda}_{11}, \hat{\lambda}_{12}, \hat{\pi}_1, \hat{\lambda}_{21}, \hat{\lambda}_{22}, \hat{\pi}_2)'$ has an asymptotic multivariate normal distribution with asymptotic mean $\boldsymbol{\gamma}$ and asymptotic covariance matrix

$$\begin{aligned} \boldsymbol{I}_f^{-1}(\boldsymbol{\gamma}) = \text{Diag} \Bigg[& \frac{\lambda_{11}(1-\lambda_{11})}{n_1 \pi_1}, \frac{\lambda_{12}(1-\lambda_{12})}{n_1(1-\pi_1)}, \frac{\pi_1(1-\pi_1)}{N_1}, \\ & \frac{\lambda_{21}(1-\lambda_{21})}{n_2 \pi_2}, \frac{\lambda_{22}(1-\lambda_{22})}{n_2(1-\pi_2)}, \frac{\pi_2(1-\pi_2)}{N_2} \Bigg]. \end{aligned}$$

Thus, for $i = 1, 2$, asymptotically we have

$$V(\hat{\lambda}_{i1}) = \frac{\lambda_{i1}(1-\lambda_{i1})}{n_i \pi_i},$$
$$V(\hat{\lambda}_{i2}) = \frac{\lambda_{i2}(1-\lambda_{i2})}{n_i(1-\pi_i)},$$
$$V(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{N_i},$$

and that $\hat{\lambda}_{11}, \hat{\lambda}_{12}, \hat{\pi}_1, \hat{\lambda}_{21}, \hat{\lambda}_{22}, \hat{\pi}_2$ are asymptotically mutually independent.

3.4 A Full Likelihood Naive Wald CI

We begin with constructing a naive Wald-type confidence interval for the risk ratio r. Note that $\hat{p}_i = \hat{\pi}_i\hat{\lambda}_{i1} + (1-\hat{\pi}_i)\hat{\lambda}_{i2}$ and that $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, and $\hat{\pi}_i$ are independent, $i = 1, 2$. Thus, using the delta method, we have

$$\begin{aligned}\sigma_i^2 &\equiv V(\hat{p}_i) \\ &\approx \left(\frac{\partial p_i}{\partial \lambda_{i1}}\right)^2 V(\hat{\lambda}_{i1}) + \left(\frac{\partial p_i}{\partial \lambda_{i2}}\right)^2 V(\hat{\lambda}_{i2}) + \left(\frac{\partial p_i}{\partial \pi_i}\right)^2 V(\hat{\pi}_i) \\ &= \frac{\pi_i\lambda_{i1}(1-\lambda_{i1})}{n_i} + \frac{(1-\pi_i)\lambda_{i2}(1-\lambda_{i2})}{n_i} + \frac{(\lambda_{i1}-\lambda_{i2})^2\pi_i(1-\pi_i)}{N_i}. \quad (3.13)\end{aligned}$$

The MLEs $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, and $\hat{\pi}_i$ are consistent estimators of λ_{i1}, λ_{i2}, and π_i, respectively. Because a continuous function of consistent estimators is consistent, we have that a consistent estimator of (3.13) is

$$\hat{\sigma}_i^2 = \frac{\hat{\pi}_i\hat{\lambda}_{i1}(1-\hat{\lambda}_{i1})}{n_i} + \frac{(1-\hat{\pi}_i)\hat{\lambda}_{i2}(1-\hat{\lambda}_{i2})}{n_i} + \frac{(\hat{\lambda}_{i1}-\hat{\lambda}_{i2})^2\hat{\pi}_i(1-\hat{\pi}_i)}{N_i}. \quad (3.14)$$

Recall that the MLE of r is $\hat{r} = \hat{p}_1/\hat{p}_2$. Again using the delta method, we have

$$\begin{aligned}\sigma_r^2 \equiv V(\hat{r}) &\approx \left(\frac{\partial r}{\partial p_1}\right)^2 V(\hat{p}_1) + \left(\frac{\partial r}{\partial p_2}\right)^2 V(\hat{p}_2) \\ &= \frac{\sigma_1^2}{p_2^2} + \frac{p_1^2\sigma_2^2}{p_2^4}, \quad (3.15)\end{aligned}$$

and a consistent estimator of (3.15) is

$$\hat{\sigma}_r^2 = \frac{\hat{\sigma}_1^2}{\hat{p}_2^2} + \frac{\hat{p}_1^2\hat{\sigma}_2^2}{\hat{p}_2^4}. \quad (3.16)$$

Therefore, an approximate $100(1-\alpha)\%$ naive Wald (nWald) CI for r is

$$\hat{r} \pm Z_{\alpha/2}\hat{\sigma}_r, \tag{3.17}$$

where $Z_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of the standard normal distribution. This interval estimator is referred to as a naive Wald CI because it results from a naive application of the Wald interval estimation method. We remark that the lower limit of the CI can be negative, especially when sample sizes are small and r is close to zero. In the case where the lower limit of the CI is negative, we replace the lower limit by zero.

3.5 A Full Likelihood Modified Wald CI

To alleviate the problem with the nWald CI, we propose a modified Wald (mWald) CI by first constructing an approximate $100(1-\alpha)\%$ CI for $\tau = \log r$. Then, we exponentiate this CI to obtain an approximate $100(1-\alpha)\%$ CI for r. Hong, Meeker and Escobar (2008) also suggested using transformation of parameters when constructing Wald-type CIs. Specifically, we let $\hat{\tau} = \log \hat{r}$. Then, using the delta method, we compute

$$\begin{aligned} \sigma_\tau^2 \equiv V(\hat{\tau}) &= V(\log \hat{p}_1 - \log \hat{p}_2) \\ &\approx \frac{V(\hat{p}_1)}{p_1^2} + \frac{V(\hat{p}_2)}{p_2^2} \\ &= \frac{\sigma_1^2}{p_1^2} + \frac{\sigma_2^2}{p_2^2}. \end{aligned} \tag{3.18}$$

Clearly, a consistent estimator of (3.18) is

$$\hat{\sigma}_\tau^2 = \frac{\hat{\sigma}_1^2}{\hat{p}_1^2} + \frac{\hat{\sigma}_2^2}{\hat{p}_2^2}. \tag{3.19}$$

Then, a $100(1-\alpha)\%$ CI for τ is

$$\hat{\tau} \pm Z_{\alpha/2}\hat{\sigma}_\tau. \tag{3.20}$$

Finally, an approximate $100(1-\alpha)\%$ mWald CI for r is obtained by exponentiating (3.20):

$$\left[\hat{r}/\exp(Z_{\alpha/2}\hat{\sigma}_\tau), \hat{r}\exp(Z_{\alpha/2}\hat{\sigma}_\tau)\right]. \tag{3.21}$$

Note that the mWald CI guarantees the lower limit of (3.21) is nonnegative.

3.6 A Full Likelihood Fieller-Type CI

We next develop a CI for r based on an interval estimation concept introduced in Fieller (1954). As noted previously, asymptotically, we have

$$\hat{p}_i \sim N(p_i, \sigma_i^2)$$

and $\hat{p}_1$ and $\hat{p}_2$ are independent. Because

$$\frac{\hat{p}_1 - r\hat{p}_2}{\sqrt{\sigma_1^2 + r^2\sigma_2^2}} \sim N(0,1)$$

is an asymptotic pivotal quantity, we can obtain an approximate $100(1-\alpha)\%$ Fieller CI by solving

$$\frac{(\hat{p}_1 - r\hat{p}_2)^2}{\hat{\sigma}_1^2 + r^2\hat{\sigma}_2^2} = Z_{\alpha/2}^2$$

for r. Let

$$\Delta \equiv \hat{p}_1^2\hat{p}_2^2 - (\hat{p}_1^2 - Z_{\alpha/2}^2\hat{\sigma}_1^2)(\hat{p}_2^2 - Z_{\alpha/2}^2\hat{\sigma}_2^2).$$

Because

$$\hat{p}_1^2 \geq \hat{p}_1^2 - Z_{\alpha/2}^2\hat{\sigma}_1^2 \text{ and } \hat{p}_2^2 \geq \hat{p}_2^2 - Z_{\alpha/2}^2\hat{\sigma}_2^2,$$

we have $\Delta \geq 0$. Moreover, $\Delta = 0$ if and only if $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = 0$. This phenomenon occurs rarely, for example, when $n_{111} = n_{211} = 0$. Clearly when $\Delta = 0$, a $100(1-\alpha)\%$ Fieller CI does not exist, which is a well-known limitation of the Fieller method. When $\Delta > 0$, an approximate $100(1-\alpha)\%$ Fieller CI for r is

$$\frac{\hat{p}_1\hat{p}_2 \pm \sqrt{\Delta}}{|\hat{p}_2^2 - Z_{\alpha/2}^2\hat{\sigma}_2^2|}.$$

4. An Example

In this section we use a real data set to compute an MLE point estimate and three CI estimates using the nWald interval, mWald interval, and Fieller interval for the risk ratio r. This dataset, displayed in Table 4, was first described in Hildesheim, Mann, Brinton, Szklo, Reeves and Rawls (1991) and was later used in Boese *et al.* (2006). The original study explored the relationship between exposure to herpes simplex virus (HSV) and invasive cervical cancer (ICC). A total of 2044 women participated in this study with 732 women in the case group and 1312 women in the control group. The western blot procedure was treated as a fallible detector of HSV. A sub-sample of the women were also tested using the

refined western blot procedure, which is a relatively accurate procedure and, thus, was treated as infallible. We regard this sub-sample as the training data in the double sampling scheme. Both false-positive and false-negative misclassification errors of HSV using the western blot procedure occurred in this study.

Table 4: Hildesheim *et al.* Data

			Fallible Classifier	
Group	Data	Infallible Classifier	0	1
Case	Training	0	13	3
		1	5	18
	Original	NA	318	375
Control	Training	0	33	11
		1	16	16
	Original	NA	701	535

NA: Not Available

In this example, we have the $p_1 = \Pr(\text{exposed to HSV} \mid \text{has ICC})$ is the probability that a patient truly has been exposed to HSV, given that she has ICC (case group), and $p_2 = \Pr(\text{exposed to HSV} \mid \text{does not have ICC})$ is the probability that a patient truly has been exposed to HSV, given that she does not have ICC (control group). Recall that $r = p_1/p_2$.

The MLE for r is $\hat{r} = 1.34$, and we give approximate 90% nWald, mWald, and Fieller CIs and their corresponding interval lengths in Table 5. For this particular example, all three interval estimators produced similar CIs. Because the lower limits of the CIs for two of the intervals (mWald and Fieller) exceed one, we conclude that statistical evidence indicates that a higher proportion of women exposed to HSV in the case group than in the control group. Thus, an association between exposure to HSV and having ICC could exist. However, the evidence for drawing this conclusion is relatively weak because the lower limits of the CIs are close to one.

Table 5: nWald, mWald, and Fieller CIs for the Hildesheim *et al.* Data

Method	CI	Length
nWald	(0.98, 1.71)	0.73
mWald	(1.02, 1.76)	0.74
Fieller	(1.02, 1.78)	0.76

5. Simulations

In this section, we describe and present the results of two Monte Carlo simulation studies to assess and compare the performance of our proposed CIs under various parameter and sample-size scenarios. The performance was evaluated in terms of CI coverage probabilities and average lengths. In particular, we considered two-sided approximate 90% CIs. Although equal sample sizes from each group were not required by these interval estimation methods, we assumed the total sample size $N_1 = N_2 = N$, training data sample size $n_1 = n_2 = n$, false-positive rate $\phi_1 = \phi_2 = \phi$, and false-negative rate $\theta_1 = \theta_2 = \theta$, to simplify the simulation studies and presentation of simulation results.

We first investigated the performance of our three proposed CI methods by varying total sample size. In this simulation, we chose the following parameter and sample-size configurations:

1. False-positive rate: $\phi = .1$,
2. False-negative rate: $\theta = .1$,
3. Ratio of the training sample size versus the total sample size: $s = n/N = 0.2$,
4. Total sample size N: from 100 to 400 with increments of 10,
5. True proportion parameters of interest (p_1, p_2): (.4, .6) and (.1, .2), corresponding to risk ratios of 2/3 and 1/2, respectively.

For each configuration of p_1, p_2, ϕ, θ, n/N, and N, we simulated $K = 10,000$ data sets. To simulate a data set, for $i = 1, 2$, we sampled $(n_{i00}, n_{i01}, n_{i10}, n_{i11})'$ using (3.1) and (X_i, Y_i) using (3.2). Then, we created the complete data $\boldsymbol{d}$ using (3.4). After a data set was created, we computed the three competing CIs for r. Once the K CIs were available for each type of CI, we computed the coverage probabilities (CPs) and the average lengths (ALs). Finally, we plotted the CPs and ALs versus sample sizes N for each type of CI.

Figures 1 and 2 display curves of CPs and ALs of the three CI estimators versus N for $(p_1, p_2) = (.4, .6)$ and $(p_1, p_2) = (.1, .2)$, respectively. When $(p_1, p_2) = (.4, .6)$, the corresponding binomial distributions are approximately symmetric about their means and, therefore, we expected the proposed CIs to perform well. Not surprisingly, Figure 1 demonstrates that both the nWald and mWald CIs had similar, close-to-nominal CPs, regardless of the sample sizes.

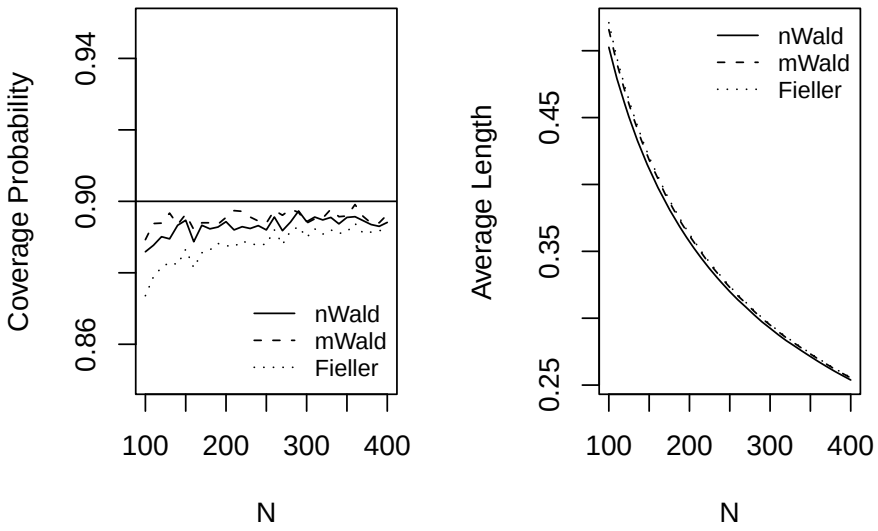

Figure 1: Coverage probabilities and average lengths versus total sample size N where $(p_1, p_2) = (.4, .6)$. The false-positive rate is $\phi = .1$, the false-negative rate is $\theta = .1$, and $s = n/N = 0.2$

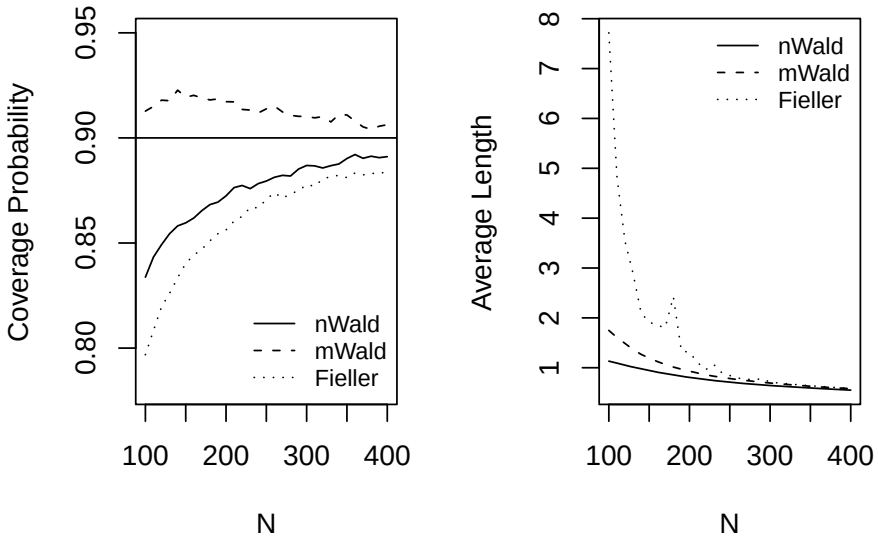

Figure 2: Coverage probabilities and average lengths versus total sample size N where $(p_1, p_2) = (.1, .2)$. The false-positive rate is $\phi = .1$, the false-negative rate is $\theta = .1$, and $s = n/N = 0.2$

The Fieller CIs had reasonable CPs for small samples ($N < 200$) and close-to-nominal CPs for large samples ($N \geq 200$). The ALs were similar for all three CIs with the nWald CIs being the narrowest and the Fieller CIs being the widest. On the other hand, when $(p_1, p_2) = (.1, .2)$, the corresponding binomial distributions were skewed and, therefore, not very well-behaved. Therefore, in this case, we did not expect the proposed CIs to perform as well for small sample sizes ($N < 200$). In fact, Figure 2 shows that both the nWald and Fieller CIs had very poor coverage for small samples ($N < 200$). However, the coverage for nWald and Fieller CIs was close to nominal when sample sizes were large ($N > 300$). Impressively, the mWald CIs had good coverage properties for all of the sample sizes considered here. For the comparison of ALs, we expected that the nWald CIs would be narrower than mWald CIs on average because naive Wald intervals commonly tend to be consistently too narrow. The Fieller CIs were generally the widest and were much wider than the other two interval estimators when the sample sizes were small. This property is very undesirable, especially with the fact that the Fieller CIs had low coverage probabilities.

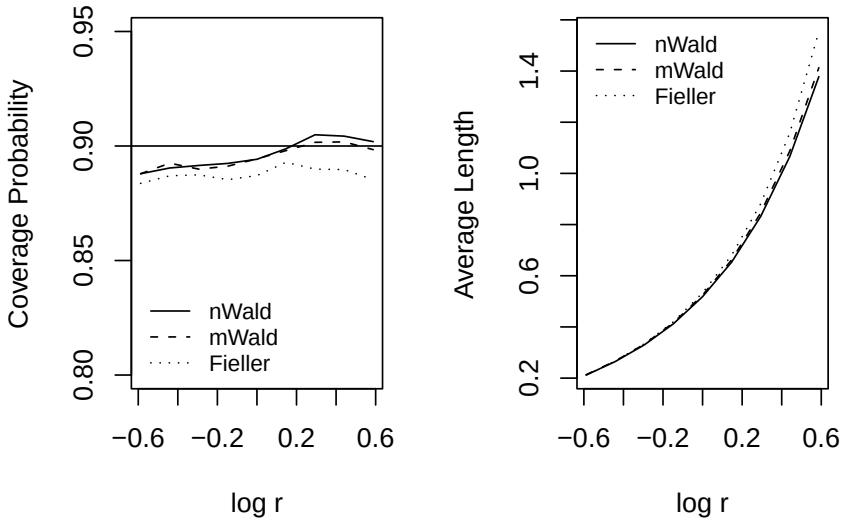

Figure 3: Coverage probabilities and average lengths versus risk ratio r where $p_1 = .5$. The false-positive rate is $\phi = .1$, the false-negative rate is $\theta = .1$, the total sample size $N = 200$, and $s = n/N = 0.2$

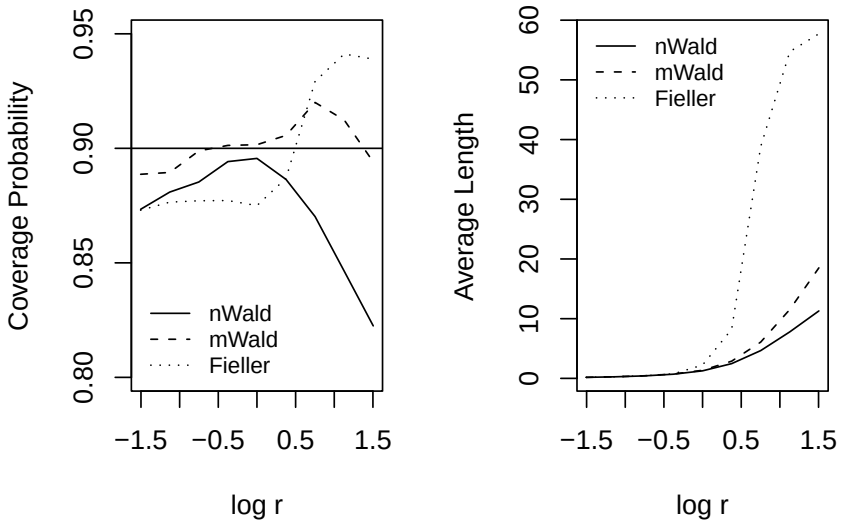

Figure 4: Coverage probabilities and average lengths versus the risk ratio r, where $p_1 = .2$. The false-positive rate is $\phi = .1$, the false-negative rate is $\theta = .1$, the total sample size $N = 200$, and $s = n/N = 0.2$

Secondly, we studied the performance of the nWald, mWald, and Fieller CIs by varying the risk ratio r. In these simulations, we chose the following parameter configurations:

1. False-positive rate: $\phi = .1$,
2. False-negative rate: $\theta = .1$,
3. Ratio of the training sample size versus the total sample size: $s = n/N = 0.2$,
4. Total sample size: $N = 200$.

We considered two simulation configurations for p_1 and p_2 with fixed values of $p_1 = .5$ and $p_1 = .2$ for the first and second simulation configurations, respectively. For each simulation configuration, we chose 9 values of p_2, $\{p_{2,1}, \cdots, p_{2,9}\}$, in an increasing order, such that $\log(r_1)$ and $\log(r_9)$ were symmetric about 0, and $\{\log(r_1), \cdots, \log(r_9)\}$ were equally spaced, where $r_t = p_1/p_{2,t}, t = 1, \cdots, 9$. We let $p_{2,9} = .9$ for both configurations. Using the assumption that $\log(r_1)$ and $\log(r_9)$ are symmetric about 0, we obtained $p_{2,1} \approx 0.278$ and $p_{2,1} \approx 0.044$ for the two configurations, respectively. Note that in this way, we ensured that the values of the parameters $\{p_{2,1}, \cdots, p_{2,9}\}$ were between 0 and 1. For each simulation

configuration, we then determined $p_{2,2}, \cdots, p_{2,8}$ such that $\{\log(r_1), \cdots, \log(r_9)\}$ were equally spaced.

For each simulation scenario with known p_1, p_2, ϕ, θ, n/N, and N, we simulated $K = 10,000$ data sets. The simulation of one data set was described previously in this section. Once the K CIs for each interval method were obtained, we calculated the coverage probabilities (CPs) and average lengths (ALs). Finally, we plotted the CPs and ALs versus r for each CI method.

Figures 3 and 4 display plots of the CPs and ALs of all CI methods versus $\log r$ for both configurations of p_1 and p_2, respectively. Figure 3 shows that both the nWald and the mWald CIs had close-to-nominal coverage for the range of $\log r$ studied here. The Fieller CI also had close-to-nominal coverage for the range of $\log r$, although the coverage was consistently slightly below the nominal level. Figure 3 also displays that the Fieller CI was slightly wider than the other two CIs. Figure 4 shows that the mWald CI had close-to-nominal coverage for the range of $\log r$ studied here. The nWald CI had close-to-nominal coverage when $\log r \in (-\log .5, \log .5)$ but much below-nominal coverage otherwise. The Fieller CI had below-nominal coverage when $\log r < .5$ and above-nominal coverage when $\log r > .5$. Figure 4 also displays that the mWald CI was slightly wider than the nWald CI. The Fieller CI was the widest and was much wider than the other two CIs when $\log r > .5$.

6. Discussion

In this article, we have considered interval estimation of the risk ratio of two binomial proportion parameters using two-sample misclassified binomial data. Because the original full likelihood function was difficult to work with, we have performed a reparameterization of the parameters. The transformed parameters in the new likelihood function were separable and, therefore, the maximum likelihood estimation was straightforward. As a result, we have derived closed-form formulas for the MLE and the nWald, the mWald, and the Fieller CIs, for the risk ratio.The nWald CI was computed using a naive application of the Wald method; the mWald CI was based on a modified Wald method that guarantees nonnegative CI limits; and the Fieller CI was constructed using an asymptotic pivotal quantity. All three CIs are easy to compute and require little computing resources.

To illustrate, all three CIs were applied to a cervical cancer data set. As expected, they produced similar CIs because the cervical cancer data have a large sample size. To compare and evaluate these three CIs, we conducted several Monte Carlo simulation studies to examine the CPs and ALs of all three CIs for r under various parameter-configuration scenarios. Because the CI estimators were developed based on asymptotic theory, we expected all three methods to perform

well for large samples. This assumption was confirmed in our simulations because the CPs were close to the nominal level for large samples and the ALs decreased as sample sizes increased.

Substantial differences in performance occurred among these three CIs. We remark that the mWald CIs had CPs close to nominal level under various parameter and sample-size scenarios. Compared with the mWald CIs, the nWald CIs were narrower but tended to have CPs less than the nominal level, especially when p_1 and p_2 were close to zero or one and the sample sizes were small ($N < 200$). The Fieller CIs generally were the widest and sometimes were much wider than the other two intervals. The behavior of the Fieller CIs was somewhat erratic because the CPs could be above or below the nominal levels, especially when p_1 and p_2 were close to zero or one and the sample sizes were small ($N < 200$). In summary, the mWald CIs consistently had nominal coverage and performed the best among three CI methods for parameter and sample-size configurations considered here and, therefore, are preferred to the nWald and Fieller intervals for the parameter and sample-size configurations considered here.

Acknowledgements

Dewi Rahardja would like to acknowledge the partial assistance of the Biostatistics Shared Resource at the Harold C. Simmons Comprehensive Cancer Center, which is supported in part by an NCI Cancer Center Support Grant, 1P30 CA142543-01. In addition, the authors thank the Editor and the referees for their thoughtful and constructive comments which have improved the presentation of this article.

References

Boese, D. H. (2003). *Likelihood-based confidence intervals for proportion parameters with binary data subject to misclassification.* PhD thesis, Baylor University.

Boese, D. H., Young, D. M. and Stamey, J. D. (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification. *Computational Statistics & Data Analysis* **50**, 3369-3385.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics* **10**, 478-486.

Evans, M., Guttman, I., Haitovsky, Y. and Swartz, T. (1996). Bayesian analysis of binary data subject to misclassification. In Berry, D., Chaloner, K. and Geweke, J., editors, *In Bayesian Analysis in Statistics and Econometrics*: *Essays in Honor of Arnold Zellner*, 67-77. John Wiley, New York.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* **16**, 175-185.

Gaba, A. and Winkler, R. L. (1992). Implications of errors in survey data: a bayesian model. *Management Science* **38**, 913-925.

Gustafson, P., Le, N. D. and Saskin, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57**, 598-609.

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

Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *Journal of American Statistical Association* **72**, 914-921.

Hong, Y., Meeker, W. and Escobar, L. (2008). Avoiding problems with normal approximation confidence intervals for probabilities. *Technometrics* **50**, 64-68.

Lee, S. C. and Byun, J. S. (2008). A bayesian approach to obtain confidence intervals for binomial proportion in a double sampling scheme subject to false-positive misclassification. *Journal of the Korean Statistical Society* **37**, 393-403.

Lie, R. T., Heuch, I. and Irgens, L. M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. *Biometrics* **50**, 433-444.

Moors, J. J. A., van der Genugten, B. B. and Strijbosch L. W. G. (2000). Repeated audit controls. *Statistica Neerlandica* **54**, 3-13.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of American Statistical Association* **65**, 1350-1361.

Tenenbein, A. (1972). A double sampling scheme for estimating from multinomial data with applications to sampling inspection. *Technometrics* **14**, 187-202.

Viana, M., Ramakrishnan, V. and Levy, P. (1993). Bayesian analysis of prevalence from results of small screening samples. *Communications Statistics*: *Theory and Methods* **22**, 575-585.

York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995). Birth defects registered by double sampling: a bayesian approach incorporating covariates and model uncertainty. *Applied Statistics* **44**, 227-242.

Received September 27, 2010; accepted December 19, 2010.

Dewi Rahardja
Department of Clinical Sciences and Simmons Cancer Center
University of Texas Southwestern Medical Center
Dallas, TX 75390-8822, USA
rahardja@gmail.com

Dean M. Young
Department of Statistical Science
Baylor University
Waco, TX 76798-7140, USA
Dean_Young@Baylor.edu