Case Study

ESTIMATION OF ODDS RATIO WITH MISCLASSIFICATION METHODS IN LUNG CANCER

Ch. Sreelatha¹, B. Muniswamy¹
¹Department of Statistics, Andhra University, Visakhapatnam-530 003, India

Correspondence should be addressed to Ch. Sreelatha

Received August 19, 2017; Accepted October 20, 2017; Published January 29, 2018;

Copyright: © 2018 Ch. Sreelatha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cite This Article: Sreelatha, C., Muniswamy, B. (2018). Estimation of odds ratio with misclassification methods in Lung cancer. BMR Medicine, 4(1). 1-4

ABSTRACT

Odds ratio is one measure used in epidemiological to study the association between the exposure and the corresponding disease. In a case control study it is conducted to examine the relationship between smoking habit and Lung cancer. In this study we work on mismeasurement method. The Mismeasurement can be separated into two types, namely non-differential and differential mismeasurement. Misclassification of exposure variables in epidemiologic studies may lead to biased estimation of parameters and loss of power in statistical inferences. Simple estimates for predictive values when misclassification is nondifferential are presented. Using them, we estimated the corrected log odds ratio. In this we have to find the effects of non-differential misclassification when 5% of smokers are misclassified as nonsmokers and 8% of non smokers are misclassified as smokers and also we are study the effects of differential misclassification when 20% of smoking and non smoking cases, but not controls, are misclassified as nonsmokers and smokers.

KEYWORDS: Odds ratio, misclassification, non-differential and differential mismeasurement, log odds ratio.

INTRODUCTION

Case-control study is still one of the most commonly used study designs in epidemiological research. Misclassification of case-control status remains a significant issue because it will bias the results of a case-control study. There exist two types of misclassification, differential versus nondifferential. It is commonly accepted that nondifferential misclassification will bias the results of the study towards the null hypothesis. Conversely, no reports have assessed the impact and direction of differential misclassification on odds ratio (OR) estimate. The goal of the present study is to demonstrate by statistical derivation that patterns exist on the bias induced by differential misclassification.

In epidemiology studies, where the assessment of the relationship between exposure and outcome variables is the main goal, misclassification of exposure variable leads to biased estimate of odds ratio. In multicenter clinical trials with an increasing number of centers, the possibility of misclassification of the exposure variable and the biases induced by it will arise. Methods to correct a possibly misclassified exposure so that the strength of the association between exposure and outcome variables can be precisely assessed have been a focus of statistical and epidemiological research for over 30 years. In 1977, Simulation analyses show that quite a number of biased odds ratios tend to move away from the null hypothesis and result in approaching zero or infinity with increasing proportion of misclassification among cases, controls, or both. These patterns are associated with the exposure status and the values of unbiased odds ratio (<1, 1, or >1).

1.1 Odds ratio

Odds ratio is one measure used in epidemiological to study the association between the exposure and the corresponding disease. Odds ratio is the ratio of the odds of disease among exposed people to the odds of disease among unexposed.
Odds represent the ratio of the probability of an event to its complement.

\[
OR = \frac{P(disease|exposed)/P(disease+free|exposed)}{P(disease|unexposed)/P(disease+free|unexposed)}
\]

The sample log odds ratio, \( \hat{\theta} \), has a less skewed distribution that is bell-shaped. Its approximating normal distribution has a mean of \( \log \theta \) and a standard error of

\[
SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}
\]

The SE decreases as the cell counts increase. Because the sampling distribution is closer to normality for \( \log \theta \) than \( \theta \), it is better to construct confidence intervals for \( \log \theta \). Transform back to form a confidence interval for \( \theta \). A large sample confidence interval for \( \log \theta \) is \( \log \theta \pm Z_{critical} \times (SE) \).

For table 1, the natural log of \( \hat{\theta} \) equals log(1.67) = 0.512, the SE of log \( \hat{\theta} \) equals 0.121. For the population, a 95% confidence interval for log \( \theta \) equals 0.512±1.96 (0.121), or (0.275, 0.749). The corresponding confidence interval for \( \theta \) is [exp(0.275), exp(0.749)] = (1.316, 2.115). The symbol \( e^\theta \), denotes the exponential function evaluated at \( x \). The exponential function is the antilog for the logarithm using the natural log scale. This means that \( e^\theta = c \) is equivalent to \( \log(c) = x \). For instance, \( e^{0.7} \) = 1.17 corresponds to \( \log(1.17) = 0.7 \). Since the confidence interval (1.316, 2.115) for \( \theta \) does not contain 1.0, the true odds of TB seem different for the two groups. We estimate that the odds of TB are at least 31% higher for subjects taking smoking than for subjects taking nonsmoker. The end points of the interval are not equally distant from \( \hat{\theta} = 1.67 \), because the sampling distribution of \( \hat{\theta} \) is skewed to the right.

### 1.3 Non-differential versus Differential Mismeasurement

Mismeasurement can be separated into two types, namely non-differential and differential mismeasurement. Non-differential mismeasurement presents if the observed exposure has no additional information about the response variable when the true value of that exposure is known. If the response variable is binary as disease or non-disease and the exposure is categorical, then the non-differential circumstance simply means possible misclassification probabilities do not vary between cases and controls. Non-differential misclassification could be caused by random errors such as fallible memory and misunderstanding questions and systematic errors such as test failures as long as errors are equally likely to occur in all levels. The perspective cohort studies by nature are more likely to relate to non-differential mismeasurement than differential mismeasurement, Because subjects are not aware of their future disease statuses and they are unable to alter their exposure values based on the unknown disease statuses. It is highly probable for patients to make the same number of errors in exposure non-differential case may turn into a
differential scenario when continuous data that are subject to measurement error are turned into categorical data.

### 1.4 Misclassification Rates

The categorical exposure is subject to misclassification, probabilities of its observed status given its true status are misclassification rates. The accuracy of a measurement tool can be reflected by misclassification rates. Misclassification rates can further help investigators correct misclassification. They can be written into a matrix form with diagonal ones representing correctly distributed probabilities and off-diagonal ones indicating misclassification rates. For a simple case with only one binary exposure and one binary outcome as variables(2*2 case), misclassification rates include four cells, which are sensitivity, 1 - sensitivity, specificity, 1 - specificity. The sensitivity of a test represents the people grouped into the exposed category given they are truly exposed. The specificity of a test indicates those are truly unexposed individuals. Sensitivity and specificity belong to classification probabilities (correctly probabilities). 1 - sensitivity is the complement of sensitivity called as false negative rate and 1 - specificity is the complement of specificity called as false positive rate. Both of them are misclassification probabilities. For non-differential misclassification, we have one pair of sensitivity and specificity as a result of same amount of misclassification cases and controls. We have two pairs of sensitivity and specificity for cases and controls separately in a differential misclassification. In sensitivity and specificity researchers sometimes need probabilities to have an inverse form of them. In this proportion truly exposed refers to positive predictive value (PPV) and the probability of unexposed refers to negative predictive value (NPV). Then sensitivity is 0.522 and specificity is 0.478.

### 1.5 Misclassification in relation to an odds ratio

Table 1.2: The effects of non-differential misclassification when 5% of smokers are misclassified as nonsmokers and 8% of non smokers are misclassified as smokers.

<table>
<thead>
<tr>
<th></th>
<th>Lung cancer</th>
<th>No Lung cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>400-20+16=396</td>
<td>300-15+20=305</td>
<td>701</td>
</tr>
<tr>
<td>Nonsmokers</td>
<td>200-16+20=204</td>
<td>250-20+15=245</td>
<td>449</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>550</td>
<td>1150</td>
</tr>
</tbody>
</table>

It gives a potential influences of a misclassified binary exposure on OR. 5 percents of smokers are misclassified into the unexposed group, and 8 percent of nonsmoking individuals are miscategorized in to the non smoking category. This kind of misclassification belongs to the non differential scenario because the same amount of misclassification is experienced by cases and controls. With the presence of misclassification OR reduces to 396*245/305*204=1.56 in contrast to 1.67(without misclassification). This suggests OR is underestimated and this underestimation is referred to Attenuation or reducing to the null.

Table 1.3: The effects of differential misclassification when 20% of smoking cases, but not controls, are misclassified as nonsmokers.

<table>
<thead>
<tr>
<th></th>
<th>Lung cancer</th>
<th>No Lung cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>400-80=320</td>
<td>300</td>
<td>620</td>
</tr>
<tr>
<td>Non smokers</td>
<td>200+80=280</td>
<td>250</td>
<td>530</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>550</td>
<td>1150</td>
</tr>
</tbody>
</table>

Table 1.4: The effects of differential misclassification when 20% of nonsmoking cases, but not controls, are misclassified as smokers.

<table>
<thead>
<tr>
<th></th>
<th>Lung cancer</th>
<th>No Lung cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>400+40=440</td>
<td>300</td>
<td>740</td>
</tr>
<tr>
<td>Non smokers</td>
<td>200+40=160</td>
<td>250</td>
<td>410</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>550</td>
<td>1150</td>
</tr>
</tbody>
</table>

Differential misclassification occurs in this case-control study. 20% of the smokers who develop Lung cancer are wrongly put into the nonsmoking group. No nonsmoking cases are misclassified into the smoker category. No misclassification happens in the control group.

From 1.3. OR becomes (320*250/300*280)=0.95 , here cases and controls as two groups receive different amounts of misclassification. Specifically controls do not have misclassification. This situation may be due to the potential influence of the recall process. From 1.4. 20 percent of the nonsmoking cases are misclassified into the exposed group and no misclassification is within the controls. Then OR turns to be 2.29, which indicates the parameter is overestimated or away from the null.

### CONCLUSION

In this study we work on mismeasurement method. In this we have to find the effects of non-differential misclassification when 5% of smokers are misclassified as nonsmokers and 8% of non smokers are misclassified as smokers and also we are study the effects of differential misclassification when 20% of smoking and non smoking cases, but not controls, are individually misclassified as nonsmokers and smokers. And also explain the misclassification rates.

### REFERENCES


