Exact inference for the risk ratio with an imperfect diagnostic test

J. REICZIGEL *, J. SINGER 2 AND ZS. LANG 1

1 University of Veterinary Medicine Budapest, Hungary
2 Accelsiors CRO & Consultancy Services Ltd, Budapest, Hungary

Received 7 February 2016; Final revision 8 August 2016; Accepted 10 August 2016; first published online 9 September 2016

SUMMARY

The risk ratio quantifies the risk of disease in a study population relative to a reference population. Standard methods of estimation and testing assume a perfect diagnostic test having sensitivity and specificity of 100%. However, this assumption typically does not hold, and this may invalidate naive estimation and testing for the risk ratio. We propose procedures that control for sensitivity and specificity of the diagnostic test, given the risks are measured by proportions, as it is in cross-sectional studies or studies with fixed follow-up times. These procedures provide an exact unconditional test and confidence interval for the true risk ratio. The methods also cover the case when sensitivity and specificity differ in the two groups (differential misclassification). The resulting test and confidence interval may be useful in epidemiological studies as well as in clinical and vaccine trials. We illustrate the method with real-life examples which demonstrate that ignoring sensitivity and specificity of the diagnostic test may lead to considerable bias in the estimated risk ratio.

Key words: Exact confidence interval, exact unconditional test, misclassification, prevalence ratio, relative risk.

INTRODUCTION

The risk ratio or relative risk (RR) quantifies the risk in a study population (say, in those exposed to a risk factor) relative to a reference population. Naive inference assumes that the applied diagnostic procedure is perfect (its sensitivity and specificity is 100%), i.e. no misclassification occurs. Unfortunately, this assumption usually does not hold, and ignoring this may result in misleading conclusions. Our aim is to present an exact unconditional test and confidence interval for RR controlling for sensitivity and specificity of the diagnostic test.

We address the situation, in which

- risks are quantified by the proportion of those having the condition (disease, recovery, etc.), which is typical in cross-sectional epidemiological studies but may also occur in other study designs;
- two independent random samples are drawn from the two populations;
- there is no classification error in the group definition (i.e. in the exposure);
- sensitivity and specificity of the diagnostic procedure are known (rather than estimated in the same or in another study).

In this case RR is defined as the ratio of two proportions, \( RR = \frac{p_2}{p_1} \), where \( p_2 \) is the proportion of diseased in the study group and \( p_1 \) is that in the reference group. Many authors use the term 'prevalence
ratio’ (PR) for this measure preserving the term ‘risk ratio’ for the incidence ratio [1–3], while others call it ‘prevalence risk ratio’ (PRR) [4, 5]. In the following text we will use the term risk ratio (RR). In cross-sectional studies and in therapeutic or vaccine trials with fixed-length follow-up this is the most natural measure to compare the groups. Guidelines for vaccine studies define vaccine efficacy as $1 – RR$, where the reference group is placebo. Despite this, in many studies odds ratios (OR) are calculated just because logistic regression has become a standard analysis tool readily available in most statistical software systems, although using OR instead of RR is repeatedly criticized by statisticians [1, 6–8].

While the impact of misclassification on the results of statistical analyses has been studied since the 1950s in the biomedical as well as in the social sciences [9–16], no exact test or confidence interval for the true PR has been proposed. In the next section we describe the proposed procedures, then we present two applications, finally we summarize the properties of the method. R code for the procedures is available at www2.univet.hu/users/jreiczig/RR_SeSp.

METHODS

Let us denote for population $i$ $(i = 1, 2)$ the true prevalence by $p_i$, the sensitivity by $Se_i$, and the specificity by $Sp_i$. Then the probability of a positive diagnosis (also called observed or apparent prevalence) in the $i$th population is $p_{ia} = p_iSe_i + (1 – p_i)(1 – Sp_i)$. This implies that taking independent samples of sizes $n_1$ and $n_2$ from the two populations, the number of test positives $x_1$ and $x_2$ follow the binomial distribution with parameters $n_1, p_{1a}$ and $n_2, p_{2a}$. Thus, the relative frequencies $x_1/n_1$ and $x_2/n_2$ are estimates of $p_{1a}$ and $p_{2a}$, therefore we will denote them by $\hat{p}_{1a}$ and $\hat{p}_{2a}$. What we will make use of in the following text is that the parameters $p_{1a}$ and $p_{2a}$ are in a one-to-one relationship with the true prevalences $p_1$ and $p_2$, i.e. a hypothesis about $p_1$ and $p_2$ can be mapped on to a corresponding hypothesis about $p_{1a}$ and $p_{2a}$, and tested using their estimates $\hat{p}_{1a}$ and $\hat{p}_{2a}$. The general equation, which describes the relationship between the parameters $p_{1a}$ and $p_{2a}$ and/or $Sp_1\neq Sp_2$ is

$p_{2a} = p_{1a}RR \frac{(Se_2 + Sp_2 – 1)}{(Se_1 + Sp_1 – 1)} – RR(1 – Sp_1) \frac{(Se_2 + Sp_2 – 1)}{(Se_1 + Sp_1 – 1)} + (1 – Sp_2). \quad (1)$

If $Se_1 = Se_2$ and $Sp_1 = Sp_2$, which can often be assumed in real-life applications, the equation simplifies to

$p_{2a} = p_{1a}RR + (1 – Sp)(1 – RR), \quad (2)$

where $Sp$ denotes the common specificity; and for a perfect diagnostic test, i.e. for $Se_1 = Se_2 = Sp_1 = Sp_2 = 1$, it reduces to

$p_{2a} = p_{1a}RR. \quad (3)$

Solving the equations (1)–(3) for RR, the following expressions are obtained:

$RR = \frac{(p_{2a} + Sp_2 – 1)(Se_1 + Sp_1 – 1)}{(p_{1a} + Sp_1 – 1)(Se_2 + Sp_2 – 1)}; \quad (4)$

$RR = (p_{2a} + Sp_1 – 1)/(p_{1a} + Sp_1 – 1), \quad (5)$

$RR = p_{2a}/p_{1a}. \quad (6)$

The point estimates for RR can be obtained by replacing $p_{1a}$ and $p_{2a}$ by $\hat{p}_{1a}$ and $\hat{p}_{2a}$ in equations (4)–(6). Note that the parameter space for the true prevalences $(p_1, p_2)$ is the unit square, whereas for $(p_{1a}, p_{2a})$ is a rectangle within the unit square, namely $[1 – Sp_1, Se_1] \times [1 – Sp_2, Se_2]$. Note also that the estimates $(\hat{p}_{1a}, \hat{p}_{2a})$ form a point in the sample space (which is actually a grid of points), rather than in the parameter space. In formula

$(\hat{p}_{1a}, \hat{p}_{2a}) \in \{0, 1/n_1, 2/n_1, \ldots, 1\} \times \{0, 1/n_2, 2/n_2, \ldots, 1\}.$

Assume now that we want to test for $H_0: RR = RR_0$, where $RR = p_{2a}/p_{1a}$ is the true risk ratio. This $H_0$ is a composite hypothesis, corresponding to a line segment in the parameter space of the true prevalences $p_1$ and $p_2$, namely the set of points in the unit square satisfying the equation $p_{2a} = p_{1a}RR_0$. If we map this onto the parameter space of the observed binomials $p_{1a}$ and $p_{2a}$ it will form another line segment. Its position depends on $RR_0$ as well as on the sensitivities and specificities according to equations (1)–(3), but it is always located within the rectangle $[1 – Sp_1, Se_1] \times [1 – Sp_2, Se_2]$. Figure 1 illustrates the position of this line segment for $H_0: RR = 2$ depending on the sensitivities $Se_1, Se_2$, and specificities $Sp_1, Sp_2$.

Testing for $H_0$ is equivalent to testing whether the parameters $p_{1a}$ and $p_{2a}$ of the observed independent binomial variables are located on the line segment corresponding to $H_0$. As this is also a composite hypothesis, it can be tested applying the intersection-union principle [17], which means that a critical (or
Inference for RR with an imperfect diagnostic test

For each element of $H_0$, i.e. for each point $(p_{1a}, p_{2a})$ on the line segment representing $H_0$, we construct a critical region (rejection region) $C_{h_0}$ in the sample space, consisting of those points, which have probability under $h_0$ less than or equal to the probability of the observed point $(\hat{p}_{1a}, \hat{p}_{2a})$. In formula

$$C_{h_0} = \{(i/n_1, j/n_2); i \in \{0, \ldots, n_1\}, j \in \{0, \ldots, n_2\}, P_{h_0}(i/n_1, j/n_2) \leq P_{h_0}(\hat{p}_{1a}, \hat{p}_{2a})\},$$

where $P_{h_0}(.)$ denotes the probability of a point (or a set of points) in the sample space, given $h_0$ is true. $C_{h_0}$ can be considered as the two-dimensional generalization of that proposed by Sterne [19] for a single binomial proportion. Let $P_{h_0}(C_{h_0})$ denote the probability of $C_{h_0}$ under $h_0$, and let $M = \max\{P_{h_0}(C_{h_0}), h_0 \in H_0\}$.

Next, for each $h_0 \in H_0$ we determine the subset $S_{h_0}$ of the sample space consisting of the points with the smallest probability under $h_0$ so that $P_{h_0}(S_{h_0}) \leq M$ but adding any further point to $S_{h_0}$ would result in $P_{h_0}(S_{h_0}) > M$. Let $C_{H_0}$ denote the intersection of all these subsets, i.e. $C_{H_0} = \bigcap_{h_0 \in H_0} S_{h_0}$. It is easy to see that $C_{H_0}$ contains $(\hat{p}_{1a}, \hat{p}_{2a})$ on its boundary. (3) Finally, the $P$ value is defined as the highest probability of $C_{H_0}$ under $H_0$ (i.e. for all simple hypotheses $h_0$ in $H_0$). In formula, the $P$ value is equal to $\max\{P_{h_0}(C_{H_0}), h_0 \in H_0\}$.

Figure 2 illustrates how the resulting critical region for $H_0$: $RR = 2$ depends on sensitivities and specificities, given the two observed prevalences are $\hat{p}_{1a} = 0.575 (n_1 = 40)$ and $\hat{p}_{2a} = 0.667 (n_2 = 36)$.

Confidence intervals for the true $RR$ can be constructed by inverting the above test. That is, lower and upper confidence limits to a given confidence level $(1 – \alpha)$ are defined as the smallest and largest true $RR_0$ not rejected by the test, i.e.

$$L = \inf\{RR_0; P_{RR_0} > \alpha\} \text{ and } U = \sup\{RR_0; P_{RR_0} > \alpha\},$$

where $P_{RR_0}$ denotes the $P$ value from testing for $H_0$: $RR = RR_0$. Computationally, $L$ and $U$ are determined by increasing $RR_0$ in small steps and performing the test. Step size may depend on the required precision. In our implementation of the algorithm the default is a multiplicative increment with step size 0.001, i.e. $RR_0$ is increased or decreased as $RR_{\text{new}} = 1.001*RR_0$ or $0.999*RR_0$. Figure 3 illustrates the procedure for observed proportions $\hat{p}_{1a} = 0.575 (n_1 = 40)$, $\hat{p}_{2a} = 0.667 (n_2 = 36)$, and $Se_1 = Se_2 = 0.91, Sp_1 = Sp_2 = 0.8$.

One-tailed testing, i.e. $H_0$: $RR = RR_0$ against $H_1$: $RR > RR_0$ (or $H_1$: $RR < RR_0$) is also possible, although there are different options to define this. Perhaps the simplest method is that one side of the line representing $H_0$ (i.e. the intersection of the critical region with that half-plane) is removed from the critical region. One-sided confidence intervals (CI) can be derived by inverting this one-sided test.

APPLICATIONS

Example 1

Everhart et al. [20] studied the seroprevalence of *Helicobacter pylori* infection in adults in the United States. The analysis was carried out stratified by age and ethnic group. The infection status was determined by an IgG ELISA assay having 91% sensitivity and 96% specificity in all groups. For illustration we now compare the group of the youngest (20–29 years) and the oldest ($\geq 70$ years), in which the observed seroprevalence was 16.7% and 56.9%, respectively. For these groups, the ratio of observed prevalences...
is 56.9/16.7 = 3.41 (95% CI 3.00–3.88), whereas the correction for sensitivity and specificity results in the true PR of 4.17 (95% CI 3.58–4.96).

Example 2

Suwancharoen et al. [21] conducted a serological survey of leptospirosis in livestock in Thailand using the microscopic agglutination test (MAT) to determine serostatus of the examined animals. Five animal species were included in the study: cattle, buffaloes, pigs, sheep and goats. Infection status of each species was measured by seroprevalence, and all other species were compared to cattle as the reference group by calculating the PRs (which are same as risk ratios). In this study the MAT test was assumed to be perfect,
Table 1. Prevalence ratios reported in Suwancharoen et al. [21] and adjusted prevalence ratios assuming sensitivity = 0.76 and specificity = 0.97 (same in all groups), as reported by Cumberland et al. [22]

<table>
<thead>
<tr>
<th>Species</th>
<th>N</th>
<th>Observed prevalence</th>
<th>Unadjusted PR (95% CI)</th>
<th>True prevalence</th>
<th>Adjusted PR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle</td>
<td>9288</td>
<td>9.9%</td>
<td>3.08 (2.79–3.41)</td>
<td>37.6%</td>
<td>3.99 (3.45–4.60)</td>
</tr>
<tr>
<td>Buffaloes</td>
<td>1376</td>
<td>30.5%</td>
<td>1.09 (0.95–1.26)</td>
<td>10.7%</td>
<td>1.13 (0.90–1.42)</td>
</tr>
<tr>
<td>Pigs</td>
<td>1898</td>
<td>10.8%</td>
<td>0.47 (0.36–0.62)</td>
<td>2.3%</td>
<td>0.24 (0.06–0.49)</td>
</tr>
<tr>
<td>Sheep</td>
<td>1110</td>
<td>4.7%</td>
<td>0.80 (0.60–1.09)</td>
<td>6.8%</td>
<td>0.72 (0.37–1.17)</td>
</tr>
<tr>
<td>Goats</td>
<td>516</td>
<td>7.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PR, Prevalence ratio; CI, confidence interval.

as this test is usually regarded to be the gold standard test. However, other studies found that the sensitivity of the MAT test is far below 100% [22–25]. Cumberland et al. [22] found the sensitivity to be 30% for first acute-phase specimens, 63% for second acute-phase specimens, and 76% for convalescent specimens. At the same time specificity was 99%, 98%, and 97%, respectively. Limmathurotsakul et al. [25] estimated the sensitivity of MAT by a Bayesian analysis and found it to be 49.8%. These findings indicate the need for an adjustment of the PR estimates.

If we consider the most optimistic scenario of these, i.e. $Se = 76\%$ and $Sp = 97\%$ [22], we obtain the adjusted PRs shown in Table 1. The difference between the unadjusted and adjusted PR is far from negligible, for example in case of sheep the adjusted ratio is less than half of the unadjusted one. In case of buffaloes, adjusted and unadjusted 95% CIs do not even overlap.

### DISCUSSION

If sensitivity and/or specificity of the diagnostic test is <100%, the observed and true prevalence may differ from each other, influencing also estimation and testing of relative risk measured by the PR. We proposed an exact unconditional test and CIs for the true PR. The method can be applied even if the sensitivities and specificities differ in the two groups, for example if patients are diagnosed by different methods or some sort of differential misclassification occurs.

Taking sensitivity and specificity into account may either increase or decrease the $P$ value compared to the one obtained without considering sensitivity and specificity. For instance, consider testing for $H_0$: $RR = 2$ with observed prevalences $\hat{p}_{1a} = 0.48 (n_1 = 50)$ and $\hat{p}_{2a} = 0.62 (n_2 = 50)$. Assuming $Se_1 = Se_2 = Sp_1 = Sp_2 = 1$ results in $P = 0.0251$, for $Se_1 = Se_2 = 0.8$ and $Sp_1 = Sp_2 = 1$ the $P$ value increases to $P = 0.0474$, whereas for $Se_1 = Se_2 = 0.6$ and $Sp_1 = Sp_2 = 1$ a smaller $P$ value of 0.0155 is obtained.

It may occur that the observed data contradict the given sensitivity and specificity. Let us assume, for example, that a certain diagnostic test is known to have $Se = 0.8$ and $Sp = 1$ and using this test we observe 90 positives out of 100. This observation is very unlikely even if the true prevalence is 100%, since under these conditions the observed variable has a binomial distribution with $n = 100$, $P = 0.8$, and the probability that it is $\geq 90$ is as low as 0.0057. The same problem may arise if the number of positives is much less than the expected minimum assuming the given sensitivity and specificity. In such cases one should consider the possibility that sensitivity and/or specificity data are incorrect.

The proposed method can be further developed in several directions. First, one can take into account misclassification also in the group definition, i.e. in the exposure. Brenner et al. [26] investigated this for the incidence ratio in a cohort study. Going further, similar methods could be worked out for models with several predictors, of which the categorical ones may also be affected by misclassification. Some results are known on correcting the OR obtained from logistic regression [27, 28] but similar results still lack for the PR.

There has been a long debate whether the PR or the OR is more appropriate to quantify the risk in the study group relative to the control group in a study design in which both of them can be calculated [1, 7, 29, 30]. Savu et al. [8] stated that RR or PR is more intuitively interpretable than the OR. In spite of this, many studies report OR estimates, even if the design permits calculation of RR or PR. Petersen & Deddens [31] emphasized that in cross-sectional studies, in particular when the disease is not rare, it is preferable to use PR instead of OR. Anyway, it is worth noting that our method can quite easily be adapted...
to testing the OR or the risk difference (RD), as the hypotheses \( H_0: \text{OR} = \text{OR}_0 \) and \( H_0: \text{RD} = \text{RD}_0 \) also correspond to a subset of the parameter space \([0,1] \times [0,1]\).

Another direction of potential improvement of the proposed methods is to extend them to the case when sensitivity and specificity are not taken as known but estimated from other samples, which may increase the variance of the RR estimate. This is analogous to the problem solved by Lang & Reiczigel [32] for estimating prevalence.

ACKNOWLEDGEMENTS

This work was supported by the Hungarian National Research Fund (grant no. OTKA K108571) and by the Research Faculty Grant 2015 of the Szent István University, Faculty of Veterinary Science.

DECLARATION OF INTEREST

None.

REFERENCES


