Revisiting Bias in Odds Ratios

Ivo M Foppa[1,2,3,*] and Fredrick S Dahlgren[2]

[1]Battelle Memorial Institute, Atlanta, Georgia, USA

[2]Influenza Division, Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, 30333 Georgia, USA

[3]Hessisches Landesprüfungs- und Untersuchungsamt im Gesundheitswesen,Abteilung I, Wolframstraße 33, 35683 Dillenburg, Germany

[*]Corresponding Author, Hessisches Landesprüfungs- und Untersuchungsamt im Gesundheitswesen,Abteilung I, Wolframstraße 33, 35683 Dillenburg, Germany,

ivo.foppa@hlpug.hessen.de

Abstract

Ratio measures of effect, such as the odds ratio (OR), are consistent, but the presumption of their unbiasedness is founded on a false premise: The equality of the expected value of a ratio and the ratio of expected values. We show that the invalidity of this assumptions is an important source of empirical bias in ratio measures of effect, which is due to properties of the expectation of ratios of count random variables. We investigate ORs (unconfounded, no effect modification), proposing a correction that leads to "almost unbiased" estimates. We also explore ORs with covariates. We find substantial bias in OR estimates for smaller sample sizes, which can be corrected by the proposed method. Bias correction is more elusive for adjusted analyses. The notion of unbiasedness of OR for the effect of interest for smaller sample sizes is challenged.

Keywords: Odds ratio, bias, effect measure

Introduction

Ratio measures of effect are widely used in epidemiology. In particular for case-control studies of etiology or intervention effectiveness, the odds ratio (OR) is of great importance (Pearce, 1993). The true OR is defined as

$$\phi = \frac{p_1(1-p_0)}{(1-p_1)p_0}, \tag{1}$$

where p_1 and p_0 represent exposure prevalences in cases and controls, respectively. If exposure prevalence remains constant over the study period and subjects are enrolled by "incidence density sampling", the OR represents the factor by which the "exposure" multiplies the incidence rate in the unexposed (Greenland and Thomas, 1982).

The consistent maximum likelihood (ML) estimator of ϕ (Gart, 1962) is

$$\mathrm{OR} = \frac{x_1\, y_0}{x_0\, y_1}, \tag{2}$$

where x_1 and x_0 are exposed and unexposed cases and y_1 and y_0 are exposed and unexposed controls, respectively. In the following discussion we use OR to refer to the ML estimator of the true OR ϕ.

The problem

Here, we investigate bias in the OR, where bias ϵ is

$$\epsilon = \mathbb{E}(\mathrm{OR}) - \phi. \tag{3}$$

Assuming independence of x_1, x_0, y_1, y_0, $\mathbb{E}(\mathrm{OR})$ can be written as

$$\mathbb{E}(\mathrm{OR}) = \mathbb{E}(x_1)\ \mathbb{E}(y_0)\ \mathbb{E}\left({}^{1}\!/_{x_0}\right)\ \mathbb{E}\left({}^{1}\!/_{y_1}\right), \tag{4}$$

but, as neither $\mathbb{E}\left({}^{1}\!/_{x_0}\right)$ nor $\mathbb{E}\left({}^{1}\!/_{y_1}\right)$ are defined because of zero denominators (Griffin, 1992), the whole expression (4) remains undefined. With the expectation undefined, bias (3) cannot be determined. If, in the context of an observational study, instances where there are either no unexposed cases (x_0) or exposed controls (y_1) will be discarded because no OR (2) can be computed. On average, the OR will therefore be better characterized by a situation where the variables in the denominator (x_0, y_1) are assigned truncated Poisson distributions; truncation here refers to restriction of the sample space of x to $\mathbb{Z}^+$. The truncated

Poisson distribution, denoted by $\text{Poi}^*(\mu)$, has the following form (Griffin, 1992):

$$\begin{aligned} \Pr(x = k|\mu) &= \frac{\mu^k e^{-\mu}}{k!} \times \frac{1}{1 - e^{-\mu}}, \text{ for } k = 1, 2, 3, \ldots \\ &\implies x \sim \text{Poi}^*(\mu) \end{aligned} \tag{5}$$

The expected value of a random variable distributed according to a truncated Poisson is given by

$$\mathbb{E}(x) = \frac{\mu}{1 - e^{-\mu}}. \tag{6}$$

Letting x_0^* and y_1^* being distributed according to a truncated Poisson distributions parametrized by μ_0 and γ_1 respectively, a "truncated" OR arises:

$$\text{OR}_t = \mathbb{E}\left(\frac{x_1\ y_0}{x_0^*\ y_1^*}\right). \tag{7}$$

The expectation of OR_t (7) is defined, but the expectations $\mathbb{E}(1/x_0^*)$ and $\mathbb{E}(1/y_1^*)$ do not have a closed-form expression. However, using Jensen's inequality (Casella and Berger, 1990), we have

$$\begin{aligned} \mathbb{E}(\text{OR}_t) &= \mathbb{E}\left(\frac{x_1\ y_0}{x_0^*\ y_1^*}\right) \\ &\geq \frac{\mathbb{E}(x_1)\,\mathbb{E}(y_0)}{\mathbb{E}(x_0^*)\,\mathbb{E}(y_1^*)}. \end{aligned} \tag{8}$$

Therefore,

$$\mathbb{E}(\text{OR}_t) \geq \phi \times \left(1 - e^{-\mu_0}\right) \times \left(1 - e^{-\gamma_1}\right), \tag{9}$$

indicating that the lower bound of $\mathbb{E}(\text{OR}_t)$ is $\phi \times (1 - e^{-\mu_0}) \times (1 - e^{-\gamma_1})$. However, as $(1 - e^{-\mu_0}) \times (1 - e^{-\gamma_1}) < 1$, it might happen that that expression neutralizes any biases.

"Almost unbiased" estimators of μ_1/μ_0 and γ_0/γ_1

An "almost unbiased" estimator (Chapman, 1952) for ratios of Poisson parameters such as μ_1/μ_0 or γ_0/γ_1, however, does exist. Chapman showed that x_1/w_0, for $w_0 = x_0 + 1$ is "almost unbiased" for μ_1/μ_0, as long as μ_0 is "not too small." The same holds for y_0/z_1, for $z_1 = y_1 + 1$ which is "almost unbiased" for γ_0/γ_1.

The expectation of the ratio x_1/w_0 can be derived as follows:

$$\begin{aligned}
\mathbb{E}(x_1/w_0) &= \mathbb{E}(x_1)\ \mathbb{E}(1/w_0) \\
&= \mu_1 \sum_{x_0=0}^{\infty} \frac{1}{w_0} \frac{\mu_0^{x_0}\, e^{-\mu_0}}{x_0!} \\
&= \mu_1 \sum_{w_0=1}^{\infty} \frac{1}{w_0} \frac{\mu_0^{w_0-1}\, e^{-\mu_0}}{(w_0-1)!} \\
&= \frac{\mu_1}{\mu_0}\, e^{-\mu_0} \sum_{w_0=1}^{\infty} \frac{\mu_0^{w_0}}{w_0!} \\
&= \frac{\mu_1}{\mu_0}\, e^{-\mu_0}\ (e^{\mu_0}-1) \\
&= \frac{\mu_1}{\mu_0}\ \left(1-e^{-\mu_0}\right)
\end{aligned} \tag{10}$$

The derivation of the expectation of the ratio y_0/z_1 follows (10). It is worth noting that the expectation of x_1/w_0 is therefore not simply $\frac{\mu_1}{\mu_0+1}$, as one might naïvely expect with $\mathbb{E}(w_0) = \mu_0 + 1$, but a negatively biased expression that quickly converges to $\frac{\mu_1}{\mu_0}$ with increasing μ_0. The equivalent, of course, holds for the expectation of y_0/z_1 Using this,

$$\mathrm{OR}_{+1} = \frac{x_1\ y_0}{w_0\ z_1}. \tag{11}$$

The expected value of OR_{+1} is

$$\begin{aligned}
\mathbb{E}(\mathrm{OR}_{+1}) &= \frac{\mu_1\ \gamma_0}{\mu_0\ \gamma_1}\ \left(1-e^{-\mu_0}\right)\ \left(1-e^{-\gamma_1}\right) \\
&= \phi\ \left(1-e^{-\mu_0}\right)\ \left(1-e^{-\gamma_1}\right).
\end{aligned} \tag{12}$$

Therefore, the expected value $\mathbb{E}(\mathrm{OR}_{+1})$ is the lower bound of $\mathbb{E}(\mathrm{OR}_t)$ (9). In contrast, Hauck et al. recommended to add 0.25 to each of the terms (x_1, x_0, y_1, y_0) of the ML estimator to calculate $\mathrm{OR}_{+0.25}$ (Hauck, Anderson, and Leahy III, 1982).

A simulation study

Unconfounded odds ratio

We simulated 100,000 data sets from case-control studies. The number of unexposed cases were simulated to arise according to a Poisson distribution with parameter $\mu_0 \in (5, 10, 20, 50, 100, 1000)$ with a true incidence rate ratio $\phi = 2$ and a control-case ratio (ratio of the *expected* number of controls to the *expected* number of cases) of 2, corresponding to expected sample sizes of 30, 60, 120, 300, 600 and 6000, respectively. ORs could not be computed for 834 and 6 datasets with $\mu_0 = 5$ and $\mu_0 = 10$, respectively, because of zero denominators. For $\mu_0 = 5$, corresponding to an expected sample size of 30, the average OR was 3.15, while the corrected analysis, that could make use of all datasets OR_{+1} was minimally biased downward, with the MSE a little less than a quarter of the one of the uncorrected ORs (Table 1).

For $\mu_0 = 5$, $\text{OR}_{+0.25}$ was even more strongly upwards biased than OR and for other values of μ_0 only marginally superior to OR, both in terms of bias and in terms of MSE. While both bias and MSE of OR and $\text{OR}_{+0.25}$ were always larger than for OR_{+1}, the differences vanished with large μ_0 (Table 1). We did not consider $\text{OR}_{+0.25}$ any further.

The odds ratio adjusted by one confounder

To investigate the situation where the odds ratio is confounded by a binary covariate, which increases the risk of the outcome independently of the exposure of interest by 50% and which is moderately independently associated with the exposure of interest (confounder odds ratio of exposed vs. unexposed=1.2). We conducted logistic regression analyses, adjusting the analysis by the confounder. We analyzed the data in the native form and after applying one of three corrections:

1. Adding one to cases unexposed to the exposure of interest and adding one to exposed controls (correction #1); OR^1_{+1}
2. Adding one to cases unexposed to either the exposure of interest or the confounder and adding one to controls exposed to either (correction #2); OR^2_{+1}
3. Adding one to cases unexposed to both the exposure of interest and the confounder and adding one to controls exposed to neither (correction #3); OR^3_{+1}

We investigated six levels of μ_{00}, which represents the mean number of cases unexposed to both the exposure of interest and the confounder ($\mu_{00} \in (5, 10, 20, 50, 100, 1000)$), corresponding to expected sample sizes of 92, 184, 368, 919, 1,838 and 18,375, respectively. The assumed control-to-case ratio was 2. For each setting we conducted 100,000 simulations and calculated mean and MSE for OR, OR^1_{+1}, OR^2_{+1} and OR^3_{+1}. For $\mu_{00} \in (5, 10)$ we also computed mean and MSE after excluding 72,382 and 5,762 datasets, respectively, for which the smallest stratum size was < 5. The uncorrected OR was substantially biased upwards for sample sizes under a thousand. OR^3_{+1} was essentially unbiased (Table 2). In the restricted analysis for $\mu_{00} = 5$ the bias for OR was lower than in the unrestricted analysis, but it was the only case for which OR^3_{+1} was substantially biased, downward about the same amount as OR was biased upwards. The other corrections always led to a downward bias and were clearly inferior to OR^3_{+1}, with consistently lower MSEs.

Discussion and conclusion

We examined bias in ratio measures of effect, in particular ORs. As the expected value of an OR is undefined, the bias is not defined either. This kind of problem for ratios of Poisson random variable is well known (Griffin, 1992)—an OR is a ratio of two such ratios. However, even though the bias is

not defined for ORs, we can examine the empirical properties of ORs. In fact, ORs are consistent, but more than trivially "biased" (the quotes are owed to the fact that this is not bias in the strict sense), i.e. on average off the true value, even if sample sizes are "reasonable". This phenomenon has been largely ignored in the epidemiologic literature. Even though these empirical biases are more pronounced for small sample sizes, they are unrelated to sample size problems of large-sample statistical methods. We have shown that empirical ratio measure biases can be improved by adding 1 to the denominators. In the absence of confounders and effect modifiers that adjustment (OR_{+1}; see equation (11)) leads to an "almost unbiased" OR estimate. We also found that the correction proposed by Gart (Gart, 1962), adding 0.25 to each count used to calculate the OR, performs poorly.

For the situation where there is one additional covariate we were able to identify a data correction procedure that works well (OR^3_{+1}). Future research is needed to better characterize the problem for more complex multivariate situations.

In summary, we examined statistical properties of the expectation of ratios of count random variables as an important source of empirical bias in ORs. This challenges the notion that ORs are, under very general assumptions, "good" estimates for the effects of interest even if sample sizes are relatively small.

Acknowledgments

The authors do not have a conflict of interest.

References

Casella, G. and R.L. Berger (1990). *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company.

Chapman, Douglas G (1952). "On tests and estimates for the ratio of Poisson means". In: *Annals of the Institute of Statistical Mathematics* 4.1, pp. 45–49.

Gart, John J (1962). "On the combination of relative risks". In: *Biometrics* 18.4, pp. 601–610.

Greenland, Sander and Duncan C Thomas (1982). "On the need for the rare disease assumption in case-control studies". In: *American journal of epidemiology* 116.3, pp. 547–553.

Griffin, Tralissa F (1992). "Distribution of the ratio of two poisson random variables". MA thesis.

Hauck, Walter W, Sharon Anderson, and Francis J Leahy III (1982). "Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2× 2 tables". In: *Journal of the American Statistical Association* 77.377, pp. 145–152.

Pearce, Neil (Dec. 1993). "What Does the Odds Ratio Estimate in a Case-Control Study?" In: *International Journal of Epidemiology* 22.6, pp. 1189–1192.

Tables

μ_0	OR	OR*	OR^{*2}
5	3.16 (16.173)†	1.99 (4.002)	3.30 (40.277)
10	2.47 (3.201)	2.00 (1.637)	2.42 (3.826)
20	2.20 (0.908)	2.00 (0.70)	2.18 (0.871)
50	2.07 (0.287)	2.00 (0.261)	2.07 (0.283)
100	2.04 (0.135)	2.00 (0.129)	2.03 (0.134)
1000	2.00 (0.013)	2.00 (0.012)	2.00 (0.013)

† Mean (MSE)

Table 1: Mean and mean square error (MSE) of OR, OR* (adding 1 to the number of unexposed cases and exposed controls) and OR ^{*2} (adding 0.25 to the number of unexposed cases and exposed controls), respectively, as a function of μ_0. The true OR was $\phi = 2$ and the control-to-case ratio was 2. For each value of μ_0 100,000 simulations were run.

μ_{00}	OR	OR^{*1}	OR^{*2}	OR^{*3}
5	2.31 (1.602)†	1.79 (0.742)	1.97 (0.853)	2.02 (1.068)
5‡	2.12 (0.722)	1.70 (0.407)	1.85 (0.463)	1.90 (0.545)
10	2.14 (0.538)	1.89 (0.387)	1.99 (0.416)	2.01 (0.452)
10‡	2.14 (0.538)	1.89 (0.387)	1.99 (0.416)	2.01 (0.452)
20	2.07 (0.233)	1.95 (0.199)	2.00 (0.206)	2.00 (0.214)
50	2.03 (0.085)	1.98 (0.08)	2.00 (0.082)	2.00 (0.083)
100	2.01 (0.042)	1.99 (0.041)	2.00 (0.041)	2.00 (0.041)
1000	2.00 (0.004)	2.00 (0.004)	2.00 (0.004)	2.00 (0.004)

† Mean (MSE)

‡ Analysis restricted to datasets for which all strata were ≥ 5.

Table 2: Mean and mean square error (MSE) of OR, OR^{*1}, OR^{*2} and OR^{*3} (see text) for different values of μ_{00} when one confounder is adjusted for (see text) by logistic regression analysis. The odds ratio is the exponentiated model coefficient corresponding to the exposure of interest. For $\mu_{00} \in (5, 10)$, mean and MSE were also calculated after exclusion of datasets in which the smallest stratum size in cases or controls was < 5 were discarded. The true OR was $\phi = 2$ and the control-to-case ratio was 2. For each value of μ_{00} 100,000 simulations were run.