



## Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes

Louise-Anne McNutt<sup>1</sup>, Chuntao Wu<sup>1</sup>, Xiaonan Xue<sup>2</sup>, and Jean Paul Hafner<sup>3</sup>

<sup>1</sup> Department of Epidemiology, School of Public Health, University at Albany, State University of New York, Rensselaer, NY.

<sup>2</sup> Department of Environmental Medicine, Division of Biostatistics, New York University School of Medicine, New York, NY.

<sup>3</sup> Departments of Pulmonary and General Internal Medicine, Samuel S. Stratton Department of Veterans Affairs Medical Center, Albany, NY.

Received for publication June 14, 2001; accepted for publication March 14, 2003.

Logistic regression yields an adjusted odds ratio that approximates the adjusted relative risk when disease incidence is rare (<10%), while adjusting for potential confounders. For more common outcomes, the odds ratio always overstates the relative risk, sometimes dramatically. The purpose of this paper is to discuss the incorrect application of a proposed method to estimate an adjusted relative risk from an adjusted odds ratio, which has quickly gained popularity in medical and public health research, and to describe alternative statistical methods for estimating an adjusted relative risk when the outcome is common. Hypothetical data are used to illustrate statistical methods with readily accessible computer software.

clinical trials; cohort studies; odds ratio; relative risk

The study of common outcomes is becoming more frequent in medicine and public health. Studies of symptoms, health behaviors, health care utilization, and even rare diseases in high-risk populations all have the potential to occur frequently (>10 percent) in a study population. This fact becomes an important consideration in deciding on the appropriate statistical analysis for a study. Typically, researchers use statistical methods designed for studies of rare diseases, sometimes incorrectly applied to studies of common outcomes. An example of this problem is the use of logistic regression to compute an estimated adjusted odds ratio and the subsequent interpretation of this estimate as a relative risk. This relation is approximately true when the incidence of outcome is less than 10 percent but usually not true when the outcome is more common. Although logistic regression may be correctly applied to studies of common outcomes, in public health we are often interested in estimating a relative risk (e.g., the probability of the outcome for one exposure group divided by the probability of the outcome for another exposure group (referent)), not the odds ratio, and it is this inference that becomes troublesome. In studies of common outcomes, the estimated odds ratio can, and often does, substantially overestimate the relative risk.

A method proposed by Zhang and Yu (1) to correct the adjusted odds ratio in cohort studies of common outcomes was proposed in 1998 and has gained popularity in medical and

public health research. A review of the *Journal Citation Reports* (accessed on May 15, 2001) identified 74 citations of this paper, and 56 reported studies utilized Zhang and Yu's method in the data analysis. Unfortunately, in most cases the method was incorrectly applied. By March 28, 2003, 214 scientific publications had cited Zhang and Yu's paper.

The purpose of this paper is to discuss the drawbacks of the Zhang and Yu method as applied by many researchers and briefly review alternative methods for estimating an adjusted relative risk and its confidence interval when the incidence of disease is common and confounding exists. The study designs we focus on include cohort studies and clinical trials with equal follow-up times for study subjects, and the cumulative incidence in at least one exposure or treatment group is greater than 10 percent.

We focus on methods that are compatible with statistical programs widely used in medical and public health research, including stratified analysis, Poisson regression, and the log-binomial model. Other methods to estimate confidence intervals of adjusted relative risks (e.g., delta method, bootstrap) have attractive properties (2, 3); however, user-friendly software is still developmental for these methods and not yet widely available to researchers. We focus here on the situation where effect modification (interaction with other factors) of the relative risk does not exist.

**COMPARISON OF AVAILABLE METHODS**

For the purpose of illustration, we created several hypothetical studies; each focuses on the association between a specific risk factor (*E*) and disease (*D*) and needs to be adjusted for a confounder (*C*). The data and calculated adjusted and crude measures of the relative risk for the method reviewed are shown in table 1. Additionally, we provide results from a simulation study that highlights the potential bias that may occur with the Zhang and Yu correction method (table 2).

**MODEL SELECTION: STUDYING ASSOCIATION VERSUS PREDICTION**

Rarely is there only one statistical model that adequately fits a set of data. Rather, researchers find themselves choosing among a few models that fairly summarize the information. The choice between models that adequately fit the data is based on various criteria, one of which is the research question. Relative risks are computed for studies that focus on measuring an association(s) between an exposure(s)/risk factor(s) and an outcome. Unlike predictive models where parsimony is revered, regression models for studies of association often keep several factors that may not explain large amounts of the variance in the outcome; however, these variables confound the association between exposure(s) and outcome sufficiently to warrant adjusting for them in the analysis (4, 5). Other criteria considered in model selection include the existence of influential individuals, extreme outliers, and other factors related to model fit (4).

**ZHANG AND YU'S PROPOSED METHOD**

Zhang and Yu proposed an intriguing, simple formula to convert an odds ratio provided by logistic regression to a relative risk (1):

$$RR = \frac{OR}{(1 - P_0) + (P_0 \times OR)}$$

In this formula,  $P_0$  is the incidence of the outcome in the nonexposed group, "OR" is an odds ratio from a logistic regression equation, and "RR" is an estimated relative risk. Most researchers apply this formula to the adjusted odds ratio to estimate an adjusted relative risk. Using the formula in this manner is *incorrect and will produce a biased estimate* when confounding is present. If no confounding exists, then regression analysis is not needed and simple calculations can be used to compute an estimated relative risk (6).

With logistic regression, an estimated relative risk can be computed for each covariate pattern (*i*):

$$RR_i = \frac{P(Y|E, x_{2i}, \dots, x_{ki})}{P(Y|\bar{E}, x_{2i}, \dots, x_{ki})} = \frac{1 + e^{-(\beta_0 + \beta_1 \bar{E} + \beta_2 x_{2i} + \dots + \beta_k x_{ki})}}{1 + e^{-(\beta_0 + \beta_1 E + \beta_2 x_{2i} + \dots + \beta_k x_{ki})}}$$

where *Y* is the outcome factor of interest (dependent variable), *E* is the exposure of interest, and  $x_2, \dots, x_k$  are

**TABLE 1. Comparison of methods to compute adjusted relative risks and confidence intervals for studies of common outcomes**

True RR*	Stratified, Mantel-Haenszel	Data†															
		Poisson		Log-binomial		Zhang and Yu		Logistic aOR*		Crude RR		Confounder level 1		Confounder level 2			
		aRR	95% CI	aRR	95% CI	RR	95% CI	aOR	95% CI	Crude RR	E = 1	D = 0	E = 1	D = 0	E = 1	D = 0	
0.33	aRR* 0.33	0.23, 0.48	0.33	0.23, 0.48	0.35	0.24, 0.50	0.21	0.24, 0.50	0.24	24	216	27	63	12	48	126	84
0.50	aRR* 0.50	0.37, 0.67	0.50	0.37, 0.67	0.52	0.38, 0.69	0.35	0.38, 0.69	0.35	36	204	27	63	18	42	126	84
1.00	aRR* 1.00	0.88, 1.14	1.00	0.79, 1.27	1.00	0.89, 1.12	1.00	0.77, 1.24	1.56	18	42	63	147	192	48	72	18
2.00	aRR* 2.00	1.60, 2.49	2.00	1.50, 2.67	2.31	1.93, 2.66	4.28	1.93, 2.66	2.77	24	36	42	168	192	48	36	54
3.00	aRR* 3.00	2.12, 4.24	3.00	2.04, 4.42	3.31	2.52, 4.15	5.05	2.52, 4.15	4.15	18	42	21	189	144	96	18	72

\* RR, relative risk; aOR, adjusted odds ratio; aRR, adjusted relative risk; CI, confidence interval.

† E denotes exposure; D denotes disease; /0 represents presence/absence of exposure or disease.

**TABLE 2. Simulation study comparing methods of estimating adjusted relative risk and coverage of confidence interval\***

Method	Median of adjusted relative risk	% of relative bias†	95% confidence interval coverage
Stratified, Mantel-Haenszel	1.9989	-0.05	93.80
Stratified, logit	1.9972	-0.14	93.90
Log-binomial	1.9999	-0.01	93.80
Poisson	2.0029	0.15	98.80
Zhang and Yu	2.3053	15.27	63.00
Logistic	4.3282	116.4	8.00

\* Sample size equals 500 with 50% of subjects exposed. Specified (true) adjusted relative risk is 2.00. Exposed group: prevalence of confounder, 60%; cumulative incidence of disease, 0.80 when confounder present and 0.40 when confounder absent. Nonexposed group: prevalence of confounder, 40%; cumulative incidence of disease, 0.40 when confounder present and 0.20 when confounder absent. One thousand random data sets were created, and each statistical method was applied to every data set to estimate the adjusted relative risk and its confidence interval.

† % of relative bias = [(median of adjusted relative risk estimated from 1,000 random data sets – true adjusted relative risk) / true adjusted relative risk] × 100.

confounders. Although the formula looks complicated, these probabilities are just the predicted values that statistical programs provide routinely. It should be noted that this formula cannot be used for classical case-control studies, as the intercept cannot be validly estimated.

In data from our studies on the health effects of violence, the Zhang and Yu correction, applied to the adjusted odds ratio and using the incidence among the unexposed for the entire sample, usually tends to be biased away from the null, suggesting that the strength of association is greater than is true. This bias occurs because the formula, used as one summary value, fails to take into consideration the more complex relation in the incidence of disease related to exposure for each covariate pattern. This finding also occurred in Zhang and Yu's simulation studies (1). Although the formula can be applied to specific covariate patterns, taking the ratio of the predicted probabilities is a simpler method to obtain covariate pattern-specific relative risks.

It is also important to note that, in general, if an outcome is common, then homogeneity of the odds ratio cannot coexist with homogeneity of the relative risk. It is useful to note that more than one statistical model may adequately fit the data; however, allowance for effect modification will depend on which model is selected.

The most difficult problem in estimating an adjusted relative risk for studies of common outcomes is not the point estimate (which we discuss below), but rather the confidence interval. Zhang and Yu's proposed confidence interval for the adjusted relative risk, computed by applying the above formula to the bounds on the adjusted odds ratio's confidence interval, also can be biased, leading one to believe that the relative risk estimate is more precise than is true (7). This bias occurs because the proposed calculation does not take into consideration the covariance between the estimated incidence and estimated odds ratio. Yu and Zhang note that a "trade-off between simplicity and precision" (8, p. 529) is at issue with their method; however, we believe that it is important, particularly when there are policy implications,

not to overstate precision. In the simulation study results presented in table 2, the computed 95 percent confidence interval coverage is only 63 percent (it should be 95 percent), suggesting that in some typical situations substantial misrepresentation of precision is possible.

## STRATIFIED ANALYSIS

One of the simplest and best-known techniques for calculating an adjusted relative risk is stratified analysis (9). Using stratified analysis, the relative risk between the risk factor of interest (*E*) and disease (*D*) is computed for each level of the confounder. These stratum-specific relative risks can be pooled together to create one adjusted relative risk, usually by taking a weighted average of the stratum-specific relative risks. Typically, the weights are chosen so that they are larger for strata with the most individuals and smaller for strata with fewer individuals (4).

## LOG-BINOMIAL MODEL

The log-binomial model has been proposed as a useful approach to compute an adjusted relative risk. Like logistic regression, the log-binomial model is used for the analysis of a dichotomous outcome. Both model the probability of the outcome (e.g., probability of disease given the exposure and confounders), and both assume that the error terms have a binomial distribution. The difference between the logistic model and the log-binomial model is the link between the independent variables and the probability of the outcome: In logistic regression, the logit function is used and, for the log-binomial model, the log function is used. In general, the log-binomial model produces an unbiased estimate of the adjusted relative risk. Although it has a couple of drawbacks, these appear to pose minimal restriction on its usefulness unless adjustment for many confounders is needed. First, the confidence interval for the adjusted relative risk computed may be narrower than is true (10, 11). As seen in table 2, our simula-

tion study results suggest that this bias is minor and similar to that found in stratified analysis. Similar coverage was seen in simulations of a range of relative risks and confounding patterns (data not shown). Second, in some situations, the log-binomial model does not converge to provide parameter estimates (10, 12). The lack of convergence may simply be due to software programs that have a default convergence criterion that is insufficient. This problem can be remedied by requiring additional iterations in the modeling fitting process. Another reason the model fits may not converge to the maximum likelihood estimate(s) is that the maximum likelihood estimates may lie near a boundary of the parameter space. When this occurs, the iteration can become stuck at the boundary, and a small adjustment of the interim fit away from the boundary may be needed to keep the iterations moving toward the value(s) that maximizes the likelihood.

### POISSON REGRESSION AND THE CONCEPT OF PLACING BOUNDS ON THE CONFIDENCE INTERVAL

Poisson regression is generally reserved for studies of rare diseases where patients may be followed for different lengths of time, such as cohort studies of rare outcomes conducted over many years with some patients being lost to follow-up. In contrast, unconditional logistic regression is typically utilized when every patient is followed for the same length of time or for a defined period with equal follow-up for subjects. For cohort studies where all patients have equal follow-up times, Poisson regression can be used in a similar manner as logistic regression, with a time-at-risk value specified as 1 for each subject. If the model adequately fits the data, this approach provides a correct estimate of the adjusted relative risk(s). For studies of common outcomes, Poisson regression is likely to compute a confidence interval(s) that is conservative, suggesting less precision than is true (tables 1 and 2). The reason Poisson regression produces wider confidence intervals compared with a log-binomial model and stratified analysis is that the Poisson errors are overestimates of binomial errors when the outcome is common (Poisson errors approximately equal binomial errors when the outcome (disease) is rare). As the examples in table 1 illustrate, although the confidence interval is more conservative, the actual difference compared with a stratified analysis is moderate. Conceptually, this interval can be thought of as bounding the true confidence interval.

Computer programs for the log-binomial and Poisson regression are widely available. For example, many generalized linear models' programs (e.g., PROC GENMOD in SAS; SAS Institute, Cary, North Carolina) can be used for both log-binomial and Poisson regression analysis. Checking the fit of the model can be done using standard methods.

### CROSS-SECTIONAL STUDIES

For cross-sectional studies, two common measures of association are the prevalence ratio and the prevalence odds ratio

(13). The mathematical computations for these measures are identical to the relative risk and the odds ratio, respectively. Thus, the methods presented in this paper can be utilized for cross-sectional studies; however, a temporal association between risk factors and "outcome" cannot be assessed.

### CONCLUSIONS

The use of an adjusted odds ratio to estimate an adjusted relative risk appropriate for studies of rare outcomes, however, may be misleading when the outcome is common. The overestimation may inappropriately affect clinical decision-making or policy development. Additionally, overestimation of the importance of a risk factor may lead to unintentional errors in the economic analysis of potential intervention programs or treatments. Options exist to obtain unbiased estimates of relative risks in studies of common outcomes. Two methods that have widely available user-friendly software and often are statistically appropriate (e.g., fit the data) include stratified analysis and log-binomial modeling.

### REFERENCES

1. Zhang J, Yu KF. What's relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;280:1690-1.
2. Rao CR. *Linear statistical inference and its applications*. 2nd ed. New York, NY: John Wiley & Sons, Inc, 1965.
3. Efron B, Tibshirani R. *An introduction to the bootstrap*. New York, NY: Chapman & Hall, 1993.
4. Kleinbaum DG, Kupper LL, Muller KE, et al. *Applied regression analysis and other multivariable methods*. Pacific Grove, CA: Brooks/Cole, 1998.
5. Rothman KJ, Greenland S, eds. *Modern epidemiology*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 1998.
6. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods*. New York, NY: Van Nostrand Reinhold, 1982.
7. McNutt LA, Hafner JP, Xue X. Correcting the odds ratio in cohort studies of common outcomes. (Letter). *JAMA* 1999;282:529.
8. Yu KF, Zhang J. Correcting the odds ratio in cohort studies of common outcomes (in reply). (Letter). *JAMA* 1999;282:529.
9. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719-48.
10. Skov T, Deddens J, Petersen MR, et al. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol* 1998;27:91-5.
11. Ma S, Wong CM. Estimation of prevalence proportion rates. (Letter). *Int J Epidemiol* 1999;28:175.
12. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 1986;123:174-84.
13. Lee J. Odds ratio or relative risk for cross-sectional data? *Int J Epidemiol* 1994;23:201-3.