Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies

Sander Greenland

From the Departments of Epidemiology and Statistics, University of California, Los Angeles, Los Angeles, CA.

Received for publication November 7, 2003; accepted for publication May 27, 2004.

Some recent articles have discussed biased methods for estimating risk ratios from adjusted odds ratios when the outcome is common, and the problem of setting confidence limits for risk ratios. These articles have overlooked the extensive literature on valid estimation of risks, risk ratios, and risk differences from logistic and other models, including methods that remain valid when the outcome is common, and methods for risk and rate estimation from case-control studies. The present article describes how most of these methods can be subsumed under a general formulation that also encompasses traditional standardization methods and methods for projecting the impact of partially successful interventions. Approximate variance formulas for the resulting estimates allow interval estimation; these intervals can be closely approximated by rapid simulation procedures that require only standard software functions.

absolute risk; case-control studies; clinical trials; cohort studies; logistic regression; odds ratio; relative risk; risk assessment

Abbreviation: GEE, generalized estimating equations.

Recently, McNutt et al. (1) noted a bias in a popular method by Zhang and Yu (2) for converting odds ratios to risk ratios. Both articles overlooked the extensive literature on estimating relative risks and other measures from fitted models. This literature addresses the problems they noted and provides valid methods for all study designs, including case-control studies, cohort studies, and clinical trials.

BACKGROUND

In 1989, Holland (3) proposed an “adjusted risk difference” using the same biased risk-ratio formula rediscovered by Zhang and Yu (2). Greenland and Holland (4) described the biases in that method and gave valid formulas for converting odds-ratio estimates into risk-difference estimates. There are now many model-based estimates and confidence intervals for risks (incidence proportions), rates, and their ratios and differences (5, pp. 414–415; 6–14) and for attributable fractions (15, 16). These methods can make use of input from logistic and other models, and most require no rare-disease assumption. One can also get valid confidence intervals for risk ratios by using Poisson regression with robust (generalized estimating equations (GEE) or
“sandwich”) variance estimates (10), which avoid the overly wide intervals noted elsewhere (1).

To describe the general ideas, suppose that \( r(x) \) is the risk or rate at level \( x \) of a regressor vector \( X \). \( X \) is a function of all exposures, confounders, and modifiers in the model; it may contain powers, product terms, splines, and so forth. For example, a study of cannabis smoking and lung cancer might use \( X = (\text{cannabis grams/year}, \text{pack-years cigarettes}, \text{age}, \text{age}^2, \text{female}, \text{age} \times \text{female}) \), where female = 1 for women, 0 for men. Suppose we want to compare average risks or rates when the distribution of \( X \) in a target population is \( p_1(x) \) versus \( p_0(x) \). These distributions usually correspond to everyone exposed versus everyone unexposed to some risk factor. In policy applications, \( p_1(x) \) and \( p_0(x) \) may represent the population distribution without versus with the application of some intervention program. Some \( X \) components (e.g., age, sex) may have the same distribution in \( p_1(x) \) and \( p_0(x) \); only those components affected by the exposure will differ. For example, \( p_1(x) \) could represent the existing joint distribution of cannabis smoking, cigarette smoking, age, and sex in the target, while \( p_0(x) \) could represent the same distribution of cannabis smoking, age, and sex, with zero cannabis assigned to everyone (for everyone, the first entry in \( X \) is shifted to zero, and the rest are unchanged).

The standardized (population-averaged) risk or rate under exposure or intervention \( j \) is \( R_j = \sum x p_j(x) r(x) \), where the sum is over the range of \( X \) in the target. The adjusted risk or rate ratio and difference are then \( \hat{R}_10 = R_1/R_0 \) and \( \hat{R}_{D10} = R_1 - R_0 \), respectively. The adjusted attributable fraction is \( (R_1 - R_0)/R_1 = \hat{R}_{D10}/R_1 = (RR_{10} - 1)/RR_{10} \); the population attributable fraction is the special case in which \( p_1(x) \) is the current distribution and \( p_0(x) \) is the distribution after exposure removal (15, 16). The covariate-specific \( RR \) formula given by McNutt et al. (1, p. 941) is a special case in which the distributions \( p_1(x) \) and \( p_0(x) \) are concentrated at single values \( x_1 \) and \( x_0 \) of \( X \), and the exposure differs between \( x_1 \) and \( x_0 \), but the other covariates do not. For example, to compare the risk from use of 200 g/year of cannabis with that of noneuse among males aged 50 years with 10 cigarette-pack-years of smoking, if \( X = (\text{cannabis grams/year}, \text{cigarette pack-years}, \text{age}, \text{age}^2, \text{female}, \text{age} \times \text{female}) \), one would take \( x_1 = (200, 10, 50, 502, 0, 50 \times 0) \) and \( x_0 = (0, 10, 50, 50^2, 0, 50 \times 0) \).

ESTIMATION

Model-based confidence intervals for the above quantities can be obtained from variance formulas (8–14). To avoid programming these formulas, intervals can instead be obtained by simulation or other resampling methods such as bootstrapping (16–18). If the comparison distributions \( p_1(x) \) and \( p_0(x) \) are also estimated, their estimates should also be resampled. There are many ways to make use of the resampling distribution of estimates; one should avoid naive use of the percentiles of the bootstrap distribution to set confidence limits, however (16, 17). Resampling methods allow use of any model to fit the risks or rates \( r(x) \), and they may also be applied by replacing \( r(x) \) with other outcome measures such as expected years of life lost or with clinical measures such as blood pressure, CD4 count, and so forth.

A key advantage of model-based estimates is that they do not require large numbers at each regressor level; the regressor values may even be unique to each individual (as would be expected when some covariates are continuous) (6–14). To avoid sparse-data artifacts, they do require that the numbers of cases and noncases be adequate relative to the number of model parameters, although this restriction can be reduced by using penalized estimation (shrinkage) or Bayesian methods to fit the model (19–21).

Model-based estimates of \( r(x) \) can be sensitive to influential data points and to model misspecification, but they nonetheless tend to have a smaller mean-squared error than do raw covariate-specific estimates (which become wildly unstable in sparse data) if the model fits well (22, chap. 12). When one standardizes over distributions similar to those in the data, the resulting summary estimates will be far less sensitive than the specific \( r(x) \) estimates; this robustness derives from the tendency of residual errors to average to zero over the data distribution.

EXAMPLE

Table 1 presents data on the relation of receptor level and staging to survival in a cohort of women with breast cancer (23, table 5.3), along with risk estimates from use of several methods. Let \( x = (x_1, x_2, x_3) \), with \( x_1 \) an indicator of low-receptor status and \( x_2 \) and \( x_3 \) indicators of stage II and III. The first set of estimates is the observed proportion of women in each column who died. The second set is from maximum-likelihood logistic regression using the model \( r(x) = \text{expit}(\alpha + \beta x) \), where \( \text{expit}(u) = e^u/(1 + e^u) \); the model fits well (e.g., likelihood-ratio \( p = 0.8 \) and the fitted risks are \( \text{expit}(a + xb) \), where \( a \) and \( b \) are the \( \alpha \) and \( \beta \) estimates). The third set is from a log-linear model \( r(x) = e^{a + \beta x} \) fit by binomial maximum likelihood (23); this model also fits well (e.g., \( p = 0.8 \) and the fitted risks are \( e^{a + \beta b} \)). The fourth set is from this log-linear model fit using the incorrect Poisson likelihood, as one would obtain by entering the observed column totals as person-years in a Poisson regression program (1, 10, 14).

To obtain a standardized risk ratio comparing the low and high receptor group using the total group as the standard, take \( p_1(x) \) and \( p_0(x) \) shown in the final two rows of table 1, multiply them against a row of estimated risks, and sum the results to get the \( R_1 \) and \( R_0 \) estimates. Under the log-linear model, the \( RR_{10} \) estimate simplifies to \( \exp(b_i) \), where \( b_i \) is the estimated \( x_i \) coefficient. The estimated standardized risks, ratios, and differences are \( \hat{R}_1 = 0.392, \hat{R}_0 = 0.237, RR_{10} = 1.65, \) and \( RD_{10} = 0.156 \) if the observed proportions are used; \( \hat{R}_1 = 0.401, \hat{R}_0 = 0.239, RR_{10} = 1.68, \) and \( RD_{10} = 0.162 \) if the logistic model is used; \( \hat{R}_1 = 0.371, \hat{R}_0 = 0.238, RR_{10} = 1.56, \) and \( RD_{10} = 0.133 \) if the log-linear model fit by binomial maximum likelihood is used; and \( \hat{R}_1 = 0.383, \hat{R}_0 = 0.235, RR_{10} = 1.63, \) and \( RD_{10} = 0.148 \) if log-linear Poisson regression is used. The Mantel-Haenszel estimates (5, p. 271) are \( RR_{10} = 1.62 \) and \( RD_{10} = 0.166 \). Thus, all the valid approximations yield similar results, as expected given the sample size and good fit of the models.

In contrast, the odds-ratio estimate \( \exp(b_i) \) from the logistic model is 2.51, and the Zhang-Yu risk-ratio estimate
(2) is 1.89; both overestimate the risk ratio, as expected given that the outcome is not rare (over a quarter of the patients died). Another invalid model-based adjustment predicts an expected number of exposed cases \( E \) from a model without exposure, then divides \( E \) into the observed number of exposed cases to get a standardized mortality ratio (22, sec. 4.3). This approach underestimates risk ratios (24); using a logistic model with only \( x_2 \) and \( x_3 \) yields \( E = 17.35 \) and a standardized mortality ratio of \( 23/17.35 = 1.33 \).

If the observed proportions are used, 95 percent confidence limits for \( \text{RR}_{10} \) are 1.06, 2.58 and for \( \text{RD}_{10} \) are 0.006, 0.304 (5, p. 263); if the logistic model is used, the limits for \( \text{RR}_{10} \) are 1.09, 2.57 and for \( \text{RD}_{10} \) are 0.013, 0.303 (8, 9); and, if the binomial log-linear model is used, the limits for \( \text{RR}_{10} \) are \( \exp(b_1 \pm 1.96v_1^{1/2}) = 1.05, 2.30 \), where \( v_1 \) is the estimated variance of \( b_1 \), and for \( \text{RD}_{10} \) are 0.023, 0.312 (9). Standard Poisson regression overestimates the variance of \( b_1 \), yielding limits for \( \text{RR}_{10} \) of 0.93, 2.87; nonetheless, GEE Poisson regression with the robust variance estimate (available in Stata proc xtgee and SAS proc genmod (25)) yields limits for \( \text{RR}_{10} \) of 1.07, 2.47 for \( \text{RR}_{10} \) and 0.021, 0.299 for \( \text{RD}_{10} \) using Poisson regression.

### CASE-CONTROL STUDIES

Cumulative case-control studies sample cases and controls from cohort members who do and do not get disease by the end of follow-up (5, pp. 110–111). Given a valid estimate of the crude (overall) risk \( r_c \) in the target population or of the ratio of case-control sampling fractions \( r_f \), one can estimate the covariate-specific risks in the target (and hence their differences and ratios) even if the disease is common. If the data are not sparse, one can use results from case-control modeling to estimate risks or rates and their contrasts (5, pp. 418–419; 26–32). For models (such as the logistic) in which the baseline odds is a multiplicative factor, \( \ln(r_c) \) or \( \ln(r_f) \) becomes a simple adjustment term to the model intercept (5, pp. 417–419; 26, 27); other models can be used, however (28, 31). Similar methods can be used to estimate risks from case-cohort studies, in which controls are sampled from all cohort members, not just noncases (5, pp. 417, 419; 33).

In density case-control studies, controls are sampled longitudinally from those at risk, in proportion to person-time (5, pp. 93–96). No adjustments are then needed to estimate rate ratios from the fitted logistic model (5, pp. 416–417; 34, 35), and intercept adjustments analogous to the cumulative

### TABLE 1. Data relating receptor level (low, high) and stage (I, II, III) to 5-year breast cancer mortality (23), observed and model-based estimates of average risk (incidence proportion) by receptor level and stage, and distributions for standardizing receptor-level comparisons to total-cohort stage distribution

<table>
<thead>
<tr>
<th></th>
<th>Stage I</th>
<th>Stage II</th>
<th>Stage III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Deaths</td>
<td>2</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Survivors</td>
<td>10</td>
<td>50</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>55</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Observed*</th>
<th>Logistic</th>
<th>Binomial†</th>
<th>Poisson‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>167 91</td>
<td>409 230</td>
<td>857 600</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( p_1(x) )</th>
<th>( p_0(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1(x) )</td>
<td>0.349 0</td>
<td>0.050 0</td>
</tr>
<tr>
<td>( p_0(x) )</td>
<td>0 0.349 0</td>
<td>0 0.500 0</td>
</tr>
</tbody>
</table>

* Deaths/total (nonparametric risk estimate).
† Log-linear risk model with receptor level and stage, fit by binomial maximum likelihood.
‡ Log-linear risk model fit by incorrect Poisson maximum likelihood.
§ \( p_1(x) \) puts the total cohort at a low receptor level; \( p_0(x) \) puts the total cohort at a high receptor level.
formulas can be used to estimate rates and rate differences (5, pp. 417; 27–30, 32).

If the analysis strata are small (sparse), as in matched analyses, special summary methods may be needed to estimate exposure-specific risks (36). To avoid sparse data, one often sees matching factors entered as simple terms in an unmatched analysis. Unfortunately, this strategy can produce bias if the matching factors are not ignorable and are modeled as continuous (e.g., age is entered directly despite being matched within 5-year categories), because case-control matching creates discontinuities in the sample factor-outcome relation at matching-category boundaries (37).

DISCUSSION

Traditional impact measures such as attributable fractions take $p_0(x)$ to be the current population distribution and $p_k(x)$ the distribution after complete exposure removal (5, p. 58; 13, 15, 18). These measures can be very misleading for policy projections: Feasible interventions can rarely achieve anything near complete exposure removal, may have untoward side effects (including adverse effects on quality of life or resources available for other purposes), and may affect the size of the population at risk. Hence, intelligent policy input requires consideration of the full spectrum of intervention limitations and side effects, rather than just traditional estimates (38–40). It further requires quantitative assessments of bias, as well as of random error (41–46); simulation confidence intervals are easily extended to subsume this task (16). Finally, because epidemiologic textbooks persist in erroneous claims otherwise (e.g., 47, p. 201), it is worth noting that attributable fractions do not approximate the etiologic fraction (fraction of cases caused by exposure) or the probability of causation, even if the disease is rare (48–50).

Rates are often substituted for risks when estimating impact measures. This substitution overstates impact on the study outcome when the exposure at issue strongly affects person-time at risk, as can occur when exposure affects other outcomes (5, p. 63; 51). One can reduce this problem by converting rates to risk estimates before standardizing, for example, by stratifying on follow-up time and then applying the exponential formula (5, p. 40): If the fitted rate in period $k$ is $r_k(x)$ and the length of period $k$ is $t_k$, an estimate of the risk over periods 1 through $K$ is $1 - \exp[-\sum_t r_t(x)t_t]$, where the sum is from $k = 1$ to $k = K$. One can also estimate risks from rates via survival models, which allow use of continuous time (52).

ACKNOWLEDGMENTS

The author thanks Katherine Hoggatt for helpful comments.

REFERENCES