American Journal of Epidemiology
Vol. 136, No. 4

Printed in U.S.A.

Efficiency Loss from Categorizing Quantitative Exposures into Qualitative Exposures in Case-Control Studies

Lue Ping Zhao[1,2] and Laurence N. Kolonel[1]

In the analysis of data from case-control studies, quantitative exposure variables are frequently categorized into qualitative exposure variables, such as quarters. The qualitative exposure variables may be scalar variables that take the median values of each quantile interval, or they may be vectors of indicator variables that represent each quantile interval. In a qualitative analysis, the scalar variables may be used to test the dose-response relation, while the indicator variables may be used to estimate odds ratios for each higher quantile interval versus the lowest. Qualitative analysis, implicitly and explicitly documented by many epidemiologists and biostatisticians, has several desirable advantages (including simple interpretation and robustness in the presence of a misspecified model or outlier values). In a quantitative analysis, the quantitative exposure variables may be directly regressed to test the dose-response relation, as well as to estimate odds ratios of interest. As this paper demonstrates, quantitative analysis is generally more efficient than qualitative analysis. Through a Monte Carlo simulation study, the authors estimated the loss of efficiency that results from categorizing a quantitative exposure variable by quartiles in case-control studies with a total of 200 cases and 200 controls. In the analysis of the dose-response relation, this loss is about 30% or more; the percentage may reach about 50% when the odds ratio for the fourth quartile interval versus the lowest is around 4. In estimating odds ratios, the loss of efficiency for the second, third, and fourth quartile intervals versus the lowest is around 90%, 75%, and 40%, respectively. The authors consider the pros and cons of each analytic approach, and they recommend that 1) qualitative analysis be used initially to estimate the odds ratios for each higher quantile interval versus the lowest to examine the dose-response relation and determine the appropriateness of the assumed underlying model; and 2) quantitative analysis be used to test the dose-response relation under a plausible log odds ratio model. *Am J Epidemiol* 1992;136:464–74.

case-control studies; epidemiologic methods; models, statistical; odds ratio; statistics

In the analysis used in many case-control studies, quantitative exposure variables are frequently categorized into qualitative exposure variables, such as quarters, if there are no apparent categorizations (for simplicity, the term "exposure variable" is abbreviated as "variable" throughout this paper). The qualitative variables may be scalar variables that take the median values of each quantile interval, or they may be vectors of indicator variables that represent each quantile interval. In a qualitative analysis, the

Received for publication August 19, 1991, and in final form March 2, 1992.

Abbreviations: EL%, percentage of efficiency loss; OR, odds ratio.

[1] Epidemiology Program, Cancer Research Center of Hawaii, Honolulu, HI

[2] Biostatistics Program, School of Public Health, University of Hawaii, Honolulu, HI.

Reprint requests to Dr. Lue Ping Zhao, Epidemiology Program, Cancer Research Center of Hawaii, 1236 Lauhala Street, Honolulu, HI 96813

This research was supported in part by grant PO1 CA 33619 from the National Institutes of Health.

The authors are grateful for the helpful advice of Drs. Ross Prentice (Fred Hutchinson Cancer Research Center, Seattle, WA), Stuart Lipsitz (Harvard School of Public Health, Boston, MA), Loic Le Marchand, and Marc Goodman (Cancer Research Center of Hawaii, Honolulu, HI).

scalar variables may be used to test the dose-response relation, while the indicator variables may be used to estimate odds ratios for each higher quantile interval versus the lowest. For example, in a case-control study of dietary nutrient intakes and cancer (1), let the binary outcome D denote a case if $D = 1$ and a control if $D = 0$. Also, let a quantitative variable E denote a particular nutrient intake, obtained through dietary questionnaires. The quantitative variable E is categorized into a scalar variable X, taking one of the four quantile values at 12.5, 37.5, 62.5, and 87.5 percent in the respective quartile intervals, and into a vector of three indicator variables (x_1, x_2, x_3), taking one of the four values (0, 0, 0), (1, 0, 0), (0, 1, 0), and (0, 0, 1) corresponding to the four quartile intervals. Based on these variables, the qualitative analysis regresses the outcome D on the scalar variable X to test the dose-response relation via a logistic regression $\Pr(D = 1 \mid X) = 1/[1 + \exp(-\alpha - \beta X)]$, where the intercept α is a nuisance parameter in the context of a case-control study and the parameter β quantifies the dose-response relation between D and E. To estimate the log odds ratios of three higher quartile intervals versus the first quartile interval, the qualitative analysis regresses the outcome D on the vectors of the indicator variables (x_1, x_2, x_3) via a logistic regression $\Pr(D = 1 \mid x_1, x_2, x_3) = 1/[1 + \exp(-\alpha - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3)]$, where the parameters $(\beta_1, \beta_2, \beta_3)$ are the log odds ratios for the three higher quartile intervals versus the lowest.

Such qualitative analyses prevail in epidemiologic studies for many reasons, some of which are enumerated here. First, prior to the full development of logistic regression, epidemiologists categorized quantitative variables and formed contingency tables to use several simpler statistical methods for data analysis, such as the Mantel-Haenszel method (2). Second, estimated odds ratios from qualitative analysis are simple to comprehend and to communicate. Third, although it assumes a logistic regression, qualitative analysis does not involve a strong model assumption; thus, the parameter estimates are quite robust even if the logistic shape does not fit. Fourth, in testing the dose-response relation, qualitative analysis is robust in the presence of outlier values, since it relies on categorized variables. Fifth, qualitative analysis may permit the comparison of results in the same population from studies with different data collection instruments, because quartile intervals may still be correct even when absolute values are not. Finally, it is generally assumed that qualitative variables are less contaminated by measurement error than are quantitative variables (however, this paper demonstrates that this intuition may not be true in some circumstances).

A quantitative analysis directly regresses an outcome D on a quantitative variable E via a logistic regression $\Pr(D = 1 \mid E) = 1/[1 + \exp(-\alpha_c - \beta_c E)]$, where the parameter β_c quantifies the dose-response relation between D and E. From the logistic regression, odds ratios may be estimated through a general log odds ratio (OR) model $\ln \mathrm{OR}(E) = \exp[\beta_c(E - E_r)]$, where E_r is a reference value of the quantitative variable. The estimated odds ratios for the three higher quartile intervals versus the lowest are given by $\ln \mathrm{OR}(E_j) = \beta_c(E_j - E_{12.5})$, where E_j is the jth quantile value of the quantitative variable and $j = 37.5$, 62.5, and 87.5.

Both qualitative and quantitative analyses may be used for testing the dose-response relation and for estimating odds ratios for higher quantile intervals versus the lowest. However, quantitative analysis in general involves a stronger assumption about the log odds ratio model, and thus may be less robust than qualitative analysis. On the other hand, qualitative analysis may be less efficient than quantitative analysis. The loss of efficiency resulting from use of qualitative analysis, the focus of this paper, is an important issue that one should consider in choosing between the two types of analysis. In a slightly different context, Lagakos (3) examined this loss in testing the dose-response relation, and estimated that the loss of efficiency with a quartile variable is around 22 percent; i.e., to achieve the same power as a qualitative analysis, the quantitative analysis needs 22 percent fewer sam-

ples. This estimate of the percentage of loss may be conservative, because the calculations were based on large sample (asymptotic) theory. Moreover, the calculated asymptotic relative efficiency may be conservative for alternatives far from the null hypothesis. Thus, this paper uses the Monte Carlo method to study the actual percentage of loss of efficiency resulting from use of qualitative analysis in case-control studies with a moderate sample size of 400 (200 cases and 200 controls).

METHODS

This Monte Carlo study simulates 2,000 replicates of case-control data so that the width of the expected 95 percent confidence interval for the coverage probability (defined below) is less than 2 percent. The assumption of normally distributed variables in cases and controls leads to a specific log odds ratio model. Under a specific log odds ratio model, the simulated data are quantitatively and qualitatively analyzed to test the dose-response relation as well as to estimate odds ratios for the three higher quartile intervals versus the lowest. The results of both analyses are then compared with regard to bias, coverage probability, power, and most importantly, percentage of efficiency loss.

Simulated data

In each of 2,000 replicates, 200 random quantitative variables for cases are simulated from a normal distribution with a mean μ_1 and a variance σ_1 (for simpler presentation, the notation σ_1, rather than the more conventional notation σ_1^2, is used here to denote the variance), while another 200 random variables for controls are similarly simulated with a mean μ_0 and a variance σ_0. Let E_{i1} and E_{i0}, $i = 1, 2, \ldots, 200$, denote the simulated variables in cases and controls, respectively. Throughout the simulation study, the mean and variance in controls are set to 3.48 and 0.64, which are based on the log-transformed saturated fat intake in male controls from a case-control study of thyroid cancer (1), although other choices of the mean and variance would lead to a similar conclusion. The simulation study uses the random number generator "rndn" in GAUSS (4) to generate normally distributed random variables on an IBM PS/2 (model 70; IBM, Armonk, New York).

The quantitative variables in cases and controls are transformed into scalar variables X_{i1} and X_{i0}, respectively, that take the median values in the four quartile intervals, and indicator variables $(x_{i11}, x_{i12}, x_{i13})$ and $(x_{i01}, x_{i02}, x_{i03})$, respectively, that represent the four quartile intervals. The quartile intervals are created according to the random variables E_{i0}, $i = 1, \ldots, 200$, in controls, since the choice of such a reference group does not significantly alter the efficiency (5). Thus, the four median values throughout the simulation study are approximately 2.56, 3.23, 3.73, and 4.40. The four quartile intervals are approximately $(-\infty, 2.81)$, (2.82, 3.48), (3.49, 4.16), and $(4.17, \infty)$.

Log odds ratio models

The conditional probability $\Pr(D = 1|E)$ follows a logistic regression, given that the random variables in the cases and controls have normal distributions (6). Consequently, the log odds ratio model of the variable E versus a reference value E_r is given by

$$\ln \mathrm{OR}(E) = \beta_1(E - E_r) + \beta_2(E^2 - E_r^2),$$

where the parameters $\beta_1 = (\mu_1/\sigma_1 - \mu_0/\sigma_0)$ and $\beta_2 = (1/\sigma_0 - 1/\sigma_1)/2$ are specified by means and variances of the random variables in the cases and controls. This model encompasses both log-linear and log-quadratic odds ratio models. If the two variances in the cases and controls are equal, i.e., $\sigma_1 = \sigma_0$, the above log odds ratio model reduces to a log-linear odds ratio model $\ln \mathrm{OR}(E) = \beta_1(E - E_r)$, where $\beta_1 = (\mu_1 - \mu_0)/\sigma_0$. The linear curves of the log-linear odds ratio model with $\beta_1 = 0$, 0.16, 0.31, 0.48, and 0.63 are plotted in figure 1, part a. The corresponding mean in cases equals $\mu_1 = \mu_0 + \beta_1\sigma_0$ and takes the values 3.48, 3.61, 3.73, 3.86, and 3.98.

If the two variances are not equal, the above model is a log-quadratic odds ratio

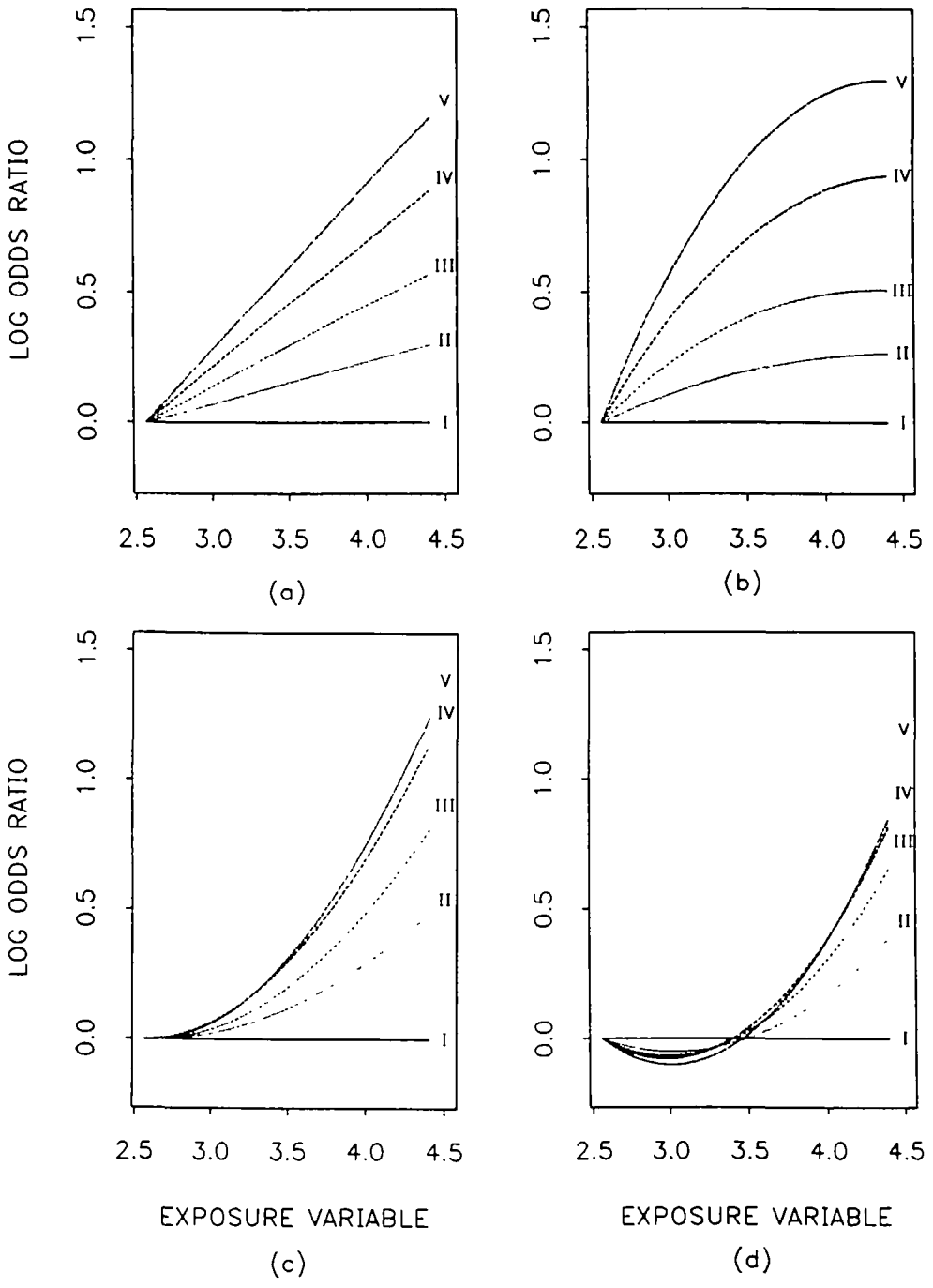

FIGURE 1. Log odds ratio models in $OR(E) = \beta_1(E - E_r) + \beta_2(E^2 - E_r^2)$ induced by normally distributed variable E. The five log odds ratio curves, labeled I, II, III, IV, and V, are plotted under the five combinations of (β_1, β_2): (0, 0), (0.16, 0), (0.31, 0), (0.48, 0), and (0.63, 0) in part a; (0, 0), (0.63, −0.07), (1.39, −0.16), (2.32, −0.26), and (3.49, −0.40) in part b; (0, 0), (−0.78, 0.15), (−1.30, 0.25), (−1.68, 0.33), and (−1.97, 0.38) in part c; and (0, 0), (−1.39, 0.23), (−2.15, 0.36), (−2.62, 0.44), and (−2.95, 0.49) in part d.

model and can also be expressed as $\ln OR(E) = \beta_2[(E - m)^2 - (E_r - m)^2]$, where $m = (\mu_1\sigma_0 - \mu_0\sigma_1)/(\sigma_0 - \sigma_1)$. This model describes three useful nonlinear curves of the dose-response relation. First, when m is close to the median of the highest quartile interval, the log odds ratio monotonically increases but the increasing rate decreases

($\beta_2 < 0$), possibly representing a plateau at higher exposure levels; the curves with m = 4.4 and μ_1 = 3.48, 3.56, 3.63, 3.71, and 3.79 are plotted in figure 1, part b. The corresponding (β_1, β_2) are (0, 0), (0.63, −0.07), (1.39, −0.16), (2.32, −0.26), and (3.49, −0.40), and the corresponding variance σ_1 in simulation takes the values 0.64, 0.58, 0.53, 0.48, and 0.42. Second, when m is close to the median of the lowest quartile interval, the log odds ratio increases with the exposure level and the increasing rate increases at higher exposure levels ($\beta_2 > 0$). The curves with m = 2.56 and μ_1 = 3.48, 3.70, 3.92, 4.14, and 4.36 are plotted in figure 1, part c. The corresponding (β_1, β_2) are (0, 0), (−0.78, 0.15), (−1.30, 0.25), (−1.68, 0.33), and (−1.97, 0.38), and the corresponding variance σ_1 in simulation takes the values 0.64, 0.79, 0.94, 1.09, and 1.24. Finally, when m is between the two median values, the log odds ratio may be U-shaped; i.e., it decreases first and then increases, indicating that a small amount of exposure is protective while an excessive amount carries a risk. The curves with m = 3.56 and μ_1 = 3.48, 3.68, 3.88, 4.08, and 4.28 are plotted in figure 1, part d. The corresponding (β_1, β_2) are (0, 0), (−1.39, 0.23), (−2.15, 0.36), (−2.62, 0.44), and (−2.95, 0.49), and the variance σ_1 in simulation takes the values 0.64, 0.90, 1.16, 1.43, and 1.69.

Bias, coverage probability, power, and percentage of efficiency loss

Three important indices for measuring the validity of a statistic are bias, coverage probability, and power, which are defined, respectively, as 1) the difference of an estimate from the true value of the parameter; 2) the probability that the derived confidence interval covers the true value; and 3) the probability that the null hypothesis is rejected when indeed it fails to hold. With 2,000 replicates in simulation, the bias may be estimated by the mean difference $\overline{\text{bias}} = (1/2{,}000)\sum_{j=1}^{2{,}000}(\hat{\beta}_j - \beta)$, where $\hat{\beta}_j$ is an estimate of the parameter β from the jth replicate. The estimate is unbiased if the mean difference is not significantly different from zero, i.e., $Z_{\text{bias}} = |\overline{\text{bias}}|/\text{SE}_\beta < 1.98$, where $|\overline{\text{bias}}|$ is the absolute value of $\overline{\text{bias}}$ and SE_β is the standard error of β_j, j = 1, . . ., 2,000; otherwise, the estimate is biased. Note that the critical value 1.98 equals the absolute value of the quantile corresponding to the upper or lower 2.5 percentile of the t distribution with 120 df, accounting for the finite sample size in the simulation study. The coverage probability (CP) is estimated by the relative frequency that the confidence interval ($\hat{\beta}_j - 1.98\text{SE}_j$, $\hat{\beta}_j + 1.98\text{SE}_j$) covers the parameter β, where SE_j is an estimated standard error. By design, the estimated coverage probability should be around 95 percent. Thus, we can regard the confidence interval as reliable if $Z_{\text{CP}} = |\text{CP} - 0.95|/\sqrt{0.05 \times 0.95} < 1.98$ and as unreliable if $Z_{\text{CP}} \geq 1.98$. Finally, the power is estimated by the frequency that the designed test statistic $Z_j = \hat{\beta}_j/\text{SE}_j$ rejects the null hypothesis H_0: $\beta = 0$, when the parameter indeed does not equal zero. Note that under the null hypothesis, the estimated power, as a complement to the coverage probability, should be around 0.05. Bias, coverage probability, and power are used to compare and assess the validity of the quantitative and qualitative analyses.

In addition, the percentage of efficiency loss (EL%), used to compare the quantitative and qualitative analyses, is defined as

$$\text{EL\%} = 100(1 - V_2/V_1),$$

where V_2 and V_1 are the sample variances of the estimates by quantitative and qualitative analysis, respectively (the ratio V_1/V_2 is often known as the asymptotic relative efficiency of the qualitative analysis versus the quantitative analysis). The EL% in general is between 0 and 100, given that qualitative analysis is less efficient than quantitative analysis. For example, a 30 percent loss of efficiency indicates that the quantitative analysis requires 30 percent fewer samples than the qualitative analysis to achieve the same power. Equivalently, to achieve the same power, the qualitative analysis requires 43 percent (100 × 30/(100 − 30)) more samples than the quantitative analysis. The higher the EL%, the greater the loss of efficiency by the qualitative analysis. Note that the EL% may be crude because of finite

sample size. However, a separate simulation shows that the estimated EL%'s are within 3–5 percent as sample size increases from 400 to 800 and to 2,000 (data not shown).

RESULTS

Table 1 shows, in two separate parts, the results for data that are simulated under the log-linear odds ratio model ln OR = $\beta_1(E - E_r)$ with β_1 = 0, 0.1, 0.3, 0.4, and 0.5. The top part of the table gives the results of testing the dose-response relation. Under the null hypothesis $\beta_1 = 0$, the biases by either type of analysis are not significantly different from zero, and the coverage probabilities are not significantly different from 95 percent. The EL% lacks direct interpretation under the null hypothesis and thus has been left blank. When the parameter β_1 differs from zero, the biases are consistently smaller for the quantitative analysis than for the qualitative analysis. In table 1, however, two biases for the quantitative analysis at β_1 = 0.31 and 0.63 are significantly different from zero, while all biases for the qualitative analysis are significantly different from zero. None of the coverage probabilities for the quantitative analysis are significantly different from 95 percent, while two of those for the qualitative analysis are significantly dif-

TABLE 1. Estimated bias, coverage probability (CP), power, and percentage of efficiency loss (EL%) by both quantitative and qualitative methods of analysis, based on simulated data under the log-linear odds ratio (OR) model ln OR(E) = $\beta_1(E - E_r)$ (figure 1, part a), for different values of β_1

	Quantitative analysis			EL%	Qualitative analysis		
	Bias†	CP	Power†		Bias	CP	Power
			Dose-response relation				
Beta							
0.00	0.0053	0.950	0.051		0.0058	0.948	0.052
0.16	0.0020	0.956	0.246	29	0.0161*	0.952	0.209
0.31	0.0066*	0.956	0.711	34	0.0380*	0.947	0.638
0.48	0.0043	0.946	0.962	39	0.0525*	0.934*	0.933
0.63	0.0115*	0.951	1.000	46	0.0788*	0.930*	0.997
			Higher quartile intervals vs. the lowest quartile interval				
ln OR							
0.00‡	0.0034	0.949	0.051		−0.0072	0.954	0.046
0.00	0.0061	0.949	0.051		−0.0085	0.952	0.048
0.00	0.0098	0.949	0.051		0.0116	0.955	0.045
0.10	0.0010	0.953	0.245	92	0.0071	0.949	0.066
0.18	0.0053	0.955	0.245	74	0.0165*	0.957	0.101
0.29	−0.0010	0.952	0.245	35	0.0290*	0.951	0.193
0.21	0.0058*	0.948	0.710	92	0.0201*	0.944	0.112
0.36	0.0139*	0.952	0.710	78	0.0236*	0.939*	0.250
0.57	0.0136*	0.955	0.710	42	0.0809*	0.947	0.610
0.31	0.0032	0.940*	0.962	92	0.0129	0.956	0.171
0.56	0.0051	0.947	0.962	76	0.0328*	0.943	0.479
0.88	0.0005	0.953	0.962	44	0.0942*	0.946	0.913
0.43	−0.0092*	0.909*	0.999	92	0.0238*	0.953	0.263
0.75	−0.0035	0.945	0.999	79	0.0373*	0.946	0.679
1.20	0.0136*	0.957	0.999	53	0.1480*	0.941	0.995

* In the "Bias" column, indicates that the bias is significantly different from zero at the 95% significance level; in the "CP" column, indicates that the coverage probability is significantly different from 95% at the 95% significance level.

† Bias, mean bias; power, power of the test statistics.

‡ The first, second, and third rows correspond to the log odds ratios for the second, third, and fourth quartile intervals vs. the first quartile interval.

ferent from 95 percent. The estimated powers are consistently higher for the quantitative analysis than for the qualitative analysis. The fifth column in the first part of the table lists the EL%, which increases from 29 percent to 46 percent as the parameter β_1 increases to 0.63. This indicates that the efficiency of the quantitative approach increases as the treatment effect moves away from zero.

The bottom portion of table 1 lists the results of estimating the log odds ratios for the three higher quartile intervals versus the lowest quartile interval. Under the null hypothesis $\beta_1 = 0$, or equivalently, the log odds ratios ln OR(E) = 0, 0, 0 for the three higher quartile intervals, none of the biases by either the quantitative analysis or the qualitative analysis are significantly different from zero, and the coverage probabilities, as complements to the estimated powers, are not significantly different from 95 percent. When the parameter β_1 increases from zero, the biases are uniformly smaller in the quantitative analysis than in the qualitative analysis. The coverage probabilities are generally not significantly different from 95 percent. The power estimates are consistently higher for the quantitative analysis than for the qualitative analysis. Most notably, the estimates for the second and third quartile intervals are much higher for the quantitative analysis than for the qualitative analysis.

The fifth column gives the EL% for the second, third, and fourth quartile intervals. For the second quartile interval, the EL% is around 92 percent regardless of variation in β; for the third quartile interval, the EL% increases from 74 percent to 79 percent as β increases; and for the fourth quartile interval, the EL% shows a substantial increase from 35 percent to 53 percent as β increases. These results suggest that the estimation of odds ratios for the qualitative analysis requires much larger samples to achieve the same power as the quantitative analysis.

Tables 2–4 give the results of analyzing data that are simulated under the log-quadratic odds ratio model ln OR(E) = $\beta_1(E - E_r) + \beta_2(E^2 - E_r^2)$, where (β_1, β_2) are chosen so that the log odds ratio ln OR(E): 1) increases at a decreasing rate (figure 1, part b); 2) monotonically increases at an increasing rate (figure 1, part c); and 3) has a U-shaped curve (figure 1, part d). Under any of these log-quadratic odds ratio models, the test statistics for the dose-response relation are not meaningful and thus are omitted.

Table 2 shows the results of analyzing data that are simulated under the log odds ratio model in figure 1, part b. Under the null hypothesis H_0: $(\beta_1, \beta_2) = (0, 0)$, the biases are consistently smaller for the quantitative analysis than for the qualitative analysis. The coverage probabilities, as complements to the estimated powers, are not significantly different from 95 percent in either analysis. When (β_1, β_2) deviate from the null, the biases are again consistently smaller in the quantitative analysis than in the qualitative analysis. Most of the biases, however, are significantly different from zero and suggest that the estimates from either the quantitative analysis or the qualitative analysis may be biased. In terms of the coverage probabilities, the quantitative analysis is worse than the qualitative analysis: Under the first two alternatives, the coverage probabilities in the quantitative analysis are all significantly different from 95 percent, while one in the qualitative analysis is significantly different from 95 percent; when the parameters are further away from the null hypothesis, the coverage probabilities in either analysis are significantly different from 95 percent.

The power estimates are consistently higher across the three quartile intervals in the quantitative analysis than in the qualitative analysis, suggesting that the former analysis is more efficient than the latter. When the log odds ratio model departs from the null, the EL% decreases from 76 percent to 60 percent, from 52 percent to 31 percent, and from 28 percent to 11 percent for the second, third, and fourth quartile intervals, respectively. This declining trend suggests that the qualitative analysis loses less efficiency under the alternatives.

Table 3 shows the results of analyzing data that are simulated under the log odds ratio

model in figure 1, part c. The results under the null hypothesis $(\beta_1, \beta_2) = (0, 0)$ by both analyses are similar to those in table 2 (omitted). As the table shows, the biases are consistently smaller for the quantitative analysis than for the qualitative analysis. The four biases for the quantitative analysis are not significantly different from zero, while all biases in the quantitative analysis are significantly different from zero. The coverage probabilities are closer to 95 percent in the qualitative analysis than in the quantitative analysis. The three coverage probabilities in the qualitative analysis are significantly different from 95 percent (probably because of the substantial amount of bias), while all of the coverage probabilities in the quantitative analysis are significantly different from 95 percent. The power estimates are consistently higher in the quantitative analysis than in the qualitative analysis. Finally, the percentage of efficiency loss increases from 82 percent to 86 percent, from 65 percent to 68 percent, and from 40 percent to 86 percent for the second, third, and fourth quartile intervals, respectively. Interestingly, the percentages for the second and third quartile intervals remain constant despite the increase in the log odds ratios, whereas the percentage for the fourth quartile interval increases as the log odds ratio increases. However, the percentages for the fourth quartile intervals under three of the alternatives may not be meaningful, because the corresponding biases of 0.2541, 0.3738, and 0.5894 in the qualitative analysis are substantial.

Table 4 provides the results of analyzing data that are simulated under the U-shaped log odds ratio model in figure 1, part d. The biases are consistently smaller in the quantitative analysis than in the qualitative analysis; most of the biases in the latter analysis are substantial. The coverage probabilities for the quantitative analysis tend to be above 95 percent, and thus the test statistic is conservative; in contrast, those for the qualitative analysis tend to be below 95 percent, and thus the test statistic is anticonservative, except when the log odds ratios are close to the null. Because of conservatism and anticonservatism in the quantitative and qualitative analyses, respectively, the power estimates falsely appear to be higher in the

TABLE 2. Results from the simulation study with the log-quadratic odds ratio (OR) model ln OR(E) = $\beta_1(E - E_r) + \beta_2(E^2 - E_r^2)$, where $\beta_1 = (\mu_0/\sigma_0 - \mu_1/\sigma_1)$, $\beta_2 = (1/\sigma_1 - 1/\sigma_0)/2$, and the log odds ratio is a convex curve that monotonically increases with E (figure 1, part b)

ln OR	Quantitative analysis			EL%†	Qualitative analysis		
	Bias†	CP†	Power†		Bias	CP	Power
0.00‡	0.0033	0.950	0.050		−0.0125	0.949	0.051
0.00	0.0018	0.953	0.048		−0.0088	0.951	0.050
0.00	−0.0054	0.951	0.050		−0.0020	0.945	0.055
0.14	0.0146*	0.934*	0.184	76	0.0187*	0.945	0.091
0.21	0.0183*	0.942	0.202	52	0.0286*	0.944	0.124
0.24	0.0128*	0.944	0.177	28	0.0248*	0.944	0.147
0.32	0.0086*	0.922*	0.579	70	0.0287*	0.953	0.200
0.48	−0.0034	0.922*	0.622	43	0.0220*	0.946	0.391
0.55	−0.0047	0.932*	0.558	20	0.0141*	0.940*	0.470
0.51	0.0297*	0.913*	0.933	67	0.0688*	0.936*	0.429
0.76	0.0353*	0.913*	0.952	40	0.0692*	0.932*	0.755
0.88	0.0369*	0.921*	0.907	19	0.0575*	0.929*	0.843
0.83	−0.0174*	0.886*	0.997	60	0.0046	0.939*	0.671
1.20	−0.0204*	0.888*	0.998	31	−0.0124	0.910*	0.938
1.40	−0.0245*	0.893*	0.993	11	−0.0447*	0.903*	0.977

* In the "Bias" column, indicates that the bias is significantly different from zero at the 95% significance level; in the "CP" column, indicates that the coverage probability is significantly different from 95% at the 95% significance level.

† Bias, mean bias; CP, coverage probability; power, power of the test statistics; EL%, percentage of efficiency loss.

‡ The first, second, and third rows correspond to the log odds ratios for the second, third, and fourth quartile intervals vs. the first quartile interval.

TABLE 3. Results from the simulation study with the log-quadratic odds ratio (OR) model ln OR(E) = $\beta_1(E - E_c) + \beta_2(E^2 - E_c^2)$, where $\beta_1 = (\mu_0/\sigma_0 - \mu_1/\sigma_1)$, $\beta_2 = (1/\sigma_1 - 1/\sigma_0)/2$, and the log odds ratio is a concave curve that monotonically increases with E (figure 1, part c)

ln OR	Quantitative analysis			EL%†	Qualitative analysis		
	Bias†	CP†	Power†		Bias	CP	Power
0.07‡	0.0060*	0.959	0.061	82	−0.0324*	0.957	0.043
0.22	−0.0004	0.964*	0.187	65	−0.0347*	0.946	0.084
0.52	−0.0067	0.966*	0.620	40	0.0821*	0.954	0.576
0.10	0.0076*	0.975*	0.100	86	−0.0397*	0.956	0.046
0.33	0.0202*	0.974*	0.449	68	−0.0266*	0.959	0.157
0.84	0.0308*	0.972*	0.990	67	0.2541*	0.900*	0.987
0.15	0.0051	0.970*	0.144	87	−0.0455*	0.951	0.059
0.44	0.0228*	0.974*	0.707	69	−0.0383*	0.957	0.213
1.12	0.0088	0.972*	1.000	77	0.3738*	0.829*	1.000
0.17	0.0165*	0.972*	0.191	86	−0.0745*	0.949	0.047
0.54	0.0096*	0.974*	0.814	68	−0.0577*	0.956	0.256
1.25	0.0812*	0.975*	1.000	86	0.5894*	0.585*	1.000

* In the "Bias" column, indicates that the bias is significantly different from zero at the 95% significance level; in the "CP" column, indicates that the coverage probability is significantly different from 95% at the 95% significance level.

† Bias, mean bias; CP, coverage probability; power, power of the test statistics; EL%, percentage of efficiency loss.

‡ The first, second, and third rows correspond to the log odds ratios for the second, third, and fourth quartile intervals vs. the first quartile interval.

qualitative analysis than in the quantitative analysis. The EL% varies from 54 percent to 95 percent, but direct interpretation is not warranted because of the substantial biases in the qualitative analysis.

DISCUSSION

The results from this Monte Carlo simulation study are consistent with Lagakos' (3) asymptotic results in that the qualitative analysis, in general, is less efficient than the quantitative analysis in testing the dose-response relation. From our simulation study, it appears that the loss of efficiency in a case-control study with a sample size of 400 is 30 percent or more. Moreover, in estimating the log odds ratios for the three higher quartile intervals versus the lowest, the qualitative analysis loses about 90 percent, 75 percent, and 52 percent efficiency, respectively. Thus, from the standpoint of efficiency, the quantitative analysis may be preferred in practice. However, the issue of the bias in the estimates must still be considered.

Bias is an important index that measures the validity of the quantitative or qualitative analysis. The results from the simulation study suggest that, under the assumption of a log odds ratio model, the biases in estimating odds ratios are consistently smaller in the quantitative analysis than in the qualitative analysis. However, these results might not hold if the assumed log odds ratio model fails. To address this issue, we simulated another five sets of case-control data under the log-quadratic odds ratio model (figure 1, part d) but used the log-linear odds ratio model to estimate the odds ratios for the quantitative and qualitative analyses. The estimated biases and coverage probabilities are listed in table 5. The biases for the second and third quartile intervals are higher for the quantitative analysis, while the biases for the fourth quartile interval are reversed. The substantially large biases in the second and third quartile intervals for the quantitative analysis are chiefly due to the misspecification of the log odds ratio model, while the lesser biases in the fourth quartile intervals for the quantitative analysis reflect the fact that the log-linear function approximates the log-quadratic odds ratio model at the two extremes. Overall, the biases in the qualitative analysis are smaller, although they are still significantly different from zero. This result suggests that the qualitative analysis

TABLE 4. Results from the simulation study with the log-quadratic odds ratio (OR) model ln OR($\dot{E}$) = $\beta_1(E - E_r) + \beta_2(E^2 - E_r^2)$, where $\beta_1 = (\mu_0/\sigma_0 - \mu_1/\sigma_1)$, $\beta_2 = (1/\sigma_1 - 1/\sigma_0)/2$, and ln OR($E$) decreases and then increases (figure 1, part d)

ln OR	Quantitative analysis			EL%†	Qualitative analysis		
	Bias†	CP†	Power†		Bias	CP	Power
−0.06‡	−0.0082*	0.977*	0.065	88	−0.108*	0.953	0.069
0.04	−0.0027	0.976*	0.026	76	−0.101*	0.945	0.044
0.36	0.0286*	0.979*	0.400	54	0.132*	0.954	0.419
−0.09	−0.0228*	0.982*	0.123	91	−0.194*	0.932*	0.122
0.08	−0.0360*	0.980*	0.024	83	−0.201*	0.920*	0.052
0.63	−0.0626*	0.980*	0.754	54	0.164*	0.955	0.853
−0.11	−0.0299*	0.980*	0.158	93	−0.257*	0.892*	0.177
0.09	−0.0368*	0.977*	0.024	87	−0.262*	0.889*	0.065
0.71	−0.0333*	0.983*	0.917	79	0.319*	0.868*	0.989
−0.15	0.0005	0.987*	0.178	95	−0.282*	0.890*	0.208
0.07	−0.0069*	0.988*	0.023	88	−0.265*	0.893*	0.066
0.78	−0.0328*	0.987*	0.972	86	0.438*	0.724*	1.000

* In the "Bias" column, indicates that the bias is significantly different from zero at the 95% significance level; in the "CP" column, indicates that the coverage probability is significantly different from 95% at the 95% significance level.

† Bias, mean bias; CP, coverage probability; power, power of the test statistics; EL%, percentage of efficiency loss.

‡ The first, second, third rows correspond to the log odds ratios for the second, third, and fourth quartile intervals vs. the first quartile interval.

TABLE 5. Results of estimating log odds ratios based on the log-linear odds ratio model using the data simulated by the log-quadratic odds ratio model (figure 1, part d)

Beta	Quantitative analysis		Qualitative analysis	
	Bias†	CP†	Bias	CP
−0.071‡	0.234	0.122	−0.144	0.950
0.029	0.260	0.504	−0.091	0.959
0.389	0.064	0.949	0.082	0.971
−0.069	0.346	0.002	−0.214	0.926
0.093	0.397	0.103	−0.213	0.910
0.619	0.148	0.891	0.178	0.947
−0.135	0.495	0.000	−0.223	0.925
0.068	0.566	0.002	−0.218	0.913
0.696	0.297	0.669	0.340	0.841
−0.169	0.587	0.000	−0.257	0.917
0.020	0.719	0.000	−0.225	0.926
0.724	0.434	0.365	0.495	0.623

† Bias, mean bias; CP, coverage probability.

‡ The first, second, and third rows correspond to the log odds ratios for the second, third, and fourth quartile intervals vs. the first quartile interval.

should be considered for consistency in estimating the log odds ratios and that these estimates may be plotted to suggest a plausible log odds ratio model. However, this result does not support the notion that the qualitative analysis always yields unbiased estimates.

The scalar variables are subject to misclassification error if the quantitative variables are themselves subject to measurement error. Consequently, a qualitative analysis with misclassified scalar variables will produce results attenuated toward the null, contradicting the common wisdom that the qualitative analysis is free from measurement error. As an illustration of this point, consider 2,000 random observations, denoted by a vector Z, from a normal distribution with a mean of 3.48 and a variance of 0.64. Suppose that a vector of observed

exposures E is unbiased for Z but is contaminated by measurement error via $E = Z + \epsilon$, where ϵ represents error from a normal distribution with mean zero and variance $0.64\lambda/(1 - \lambda)$ and where $\lambda = \text{var}(\epsilon)/[\text{var}(Z) + \text{var}(\epsilon)]$ is the percentage of variance due to the measurement error. Now both Z and E are categorized by quartiles, denoted Z_d and E_d. The correlations between E and Z and between E_d and Z_d indirectly quantify the effects due to the measurement error and the misclassification error, respectively. Under the measurement error percentages λ = 0.01, 0.10, 0.20, and 0.50, the correlations between E and Z and between E_d and Z_d are (1.00, 0.97) (0.95, 0.89), (0.89, 0.82), and (0.71, 0.64). The correlation between the two qualitative observations is worse than that between the two quantitative observations, suggesting that the scalar variables contain as much measurement error as the quantitative variables.

In conclusion, quantitative analysis under a correctly assumed model is generally more efficient and yields less biased estimates of the log odds ratios than does qualitative analysis. On the other hand, qualitative analysis may yield more reliable estimates of the odds ratios with an unspecified and possibly nonlinear log odds ratio model, but it entails a substantial loss of efficiency. In practical application, our recommendations are that 1) qualitative analysis be used to estimate the log odds ratios for the higher quantile intervals relative to the lowest quantile interval, and that these estimates be used to establish a plausible dose-response relation; and 2) quantitative analysis then be considered to test the dose-response relation. In adopting such a strategy of analysis, however, one should not select variables for the quantitative analysis based on their significance in the qualitative analysis. Such a practice could inflate the type I error, leading to invalid conclusions.

REFERENCES

1. Kolonel LN, Hankin JH, Wilkens LR, et al. An epidemiologic study of thyroid cancer in Hawaii. Cancer Causes Control 1990;1:223–34.
2. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.
3. Lagakos SW. Effects of mismodeling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 1988;7: 257–74.
4. Aptech Systems, Inc. The GAUSS system, version 2.0. Kent, WA: Aptech Systems, Inc, 1984.
5. Hsieh CC, Maisoneuve P, Boyle P, et al. Analysis of quantitative data by quantiles in epidemiological studies: classification according to cases, noncases, or all subjects? Epidemiology 1991;2:137–40.
6. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. J Am Stat Assoc 1975;70:892–8.