STATISTICS IN MEDICINE, VOL. 2, 243–251 (1983)

TESTS FOR INTERACTION IN EPIDEMIOLOGIC STUDIES: A REVIEW AND A STUDY OF POWER

SANDER GREENLAND

Division of Epidemiology, School of Public Health, University of California, Los Angeles, CA 90024, U.S.A.

SUMMARY

Tests for statistical interaction have come into increasing use in epidemiologic analysis, with most based on either an additive or multiplicative model for joint effects. Further procedures have been proposed for testing the goodness-of-fit and comparing the fit of the latter models. This paper reviews the relationships between the various tests and model comparison methods, and, for the special case of two dichotomous risk factors, presents asymptotic power functions for tests of additivity and multiplicativity. For a range of sample sizes and factor effects, the powers of the tests are computed using both the asymptotic power function and simulation studies. The powers of the tests are very low in several commonly encountered situations. In addition, convergence to the asymptotic distribution appears slow for some of the statistics. The results also indicate that likelihood comparison procedures can provide a useful adjunct to the classical hypothesis-testing approach.

KEY WORDS Case-control studies Interaction Odds ratio Risk

INTRODUCTION

Beginning about 1973, there developed an extensive discussion in the epidemiological literature concerned with the selection of the 'proper' scale (or model, or parameters) for measurement of effects of risk factors on disease incidence.[1-12] This discussion generated new research into model comparison methods and tests for statistical interaction (equivalently termed 'tests of homogeneity', 'tests for heterogeneity', or 'tests for modification' of effect parameters).[13-19] Nearly all of the research has focused on two simple types of models for joint effects: additive and multiplicative. This paper reviews the major tests and comparison methods, and examines both the asymptotic and small-sample behaviour of several of the tests and a model comparison method in the analysis of a study of two dichotomous risk factors.

There are a number of equivalent formulations for the additive and multiplicative models. We employ the following: suppose X and Y are two continuous or dichotomous risk factors for a dichotomous disease outcome D. Let $R(x, y) = \text{Prob}\ (\text{Disease} \mid X = x, Y = y)$ be the (cumulative) incidence rate of disease at level x of X and y of Y, and define the rate ratio ('relative risk') as $RR(x, y) = R(x, y)/R(x_0, y_0)$, where x_0 and y_0 are some fixed baseline or reference values for X and Y. (For example, if X represented smoking habits (in number of cigarettes per day), Y represented gender (0 = female, 1 = male), and D represented lung cancer, x_0 could represent 'non-smoker' ($x_0 = 0$) and y_0 could represent the female category ($y_0 = 0$).) A basic parameterization of the additive model for the joint effect of X and Y is

$$RR(x, y) = 1 + \alpha_1 x + \alpha_2 y \qquad \text{(model 1)}$$

and for the multiplicative model is

$$RR(x, y) = \exp(\beta_1 x + \beta_2 y). \qquad \text{(model 2)}$$

0277–6715/83/020243–09$01.00

Received September 1982

If the effect of, say, a change in X from x_1 to x_2 is measured by the amount such a change *adds* to the RR (i.e. if an additive scale is used to measure effects), then the additive model implies that this additive effect is constant across levels of Y. In model 1, this constant effect is simply $\alpha_1(x_2-x_1)$. If, on the other hand, effects are measured by the amount the change in X *multiplies* the RR (i.e. if a multiplicative scale is used to measure effects), then the multiplicative model implies that this multiplicative effect is constant across levels of Y. In model 2, this constant effect is $\exp[\beta_1(x_2-x_1)]$. If the disease incidence rate is low (say, under 0·02 or so) at all exposure levels, model 2 is well approximated by the first-order logistic model

$$OR(x, y) = \exp(\beta_1 x + \beta_2 y)$$

where $OR(x, y)$ is the disease odds ratio given by

$$OR(x, y) = \frac{R(x, y)[1-R(x_0, y_0)]}{[1-R(x, y)]R(x_0, y_0)}$$

Model 1 is strictly incompatible with model 2 if both X and Y have non-zero effects.[6, 10] Models 1 and 2 deal only with the case of two variables, but extensions to the case of three or more variables are straightforward and given in, for example, Reference 18. Models 1 and 2 may also be used with $R(x, y)$ representing person-time incidence.[18]

TESTS OF THE MODELS

Various methods have been proposed for testing and measuring the goodness-of-fit of the above models, and for comparing the fit of the models in order to choose the 'best' model for further data analysis.[6, 13, 14, 17-22] The tests are usually cast in the form of 'tests of homogeneity' (of effects) or 'tests for (statistical) interaction'.[6, 12, 22] In 'tests for interaction' the fit of each model is compared to the fit of a version of the model with an additional 'interaction' (product) term. For example, the additive model 1 might be tested against the expanded model

$$RR(x, y) = 1 + \alpha_1 x + \alpha_2 y + \alpha_3 xy \qquad \text{(model 3)}$$

Similarly, the multiplicative model 2 might be tested against the expanded model

$$RR(x, y) = \exp(\beta_1 x + \beta_2 y + \beta_3 xy) \qquad \text{(model 4)}$$

The additional parameters α_3 and β_3 are measures of departure from models 1 and 2, respectively. Such tests may be performed using likelihood-ratio (LR) statistics,[23] score (Rao) statistics,[23] or Wald (maximum-likelihood) statistics.[24]

To describe the LR tests, let $LL(k)$ be the log-likelihood of model k, where $k = 1, 2, 3, 4$. The LR test-of-fit for model 1 given model 3, a test of additivity of effects, is given by

$$G^2(\alpha_3 = 0) = -2[LL(1) - LL(3)]$$

$G^2(\alpha_3 = 0)$ is a test statistic for the hypothesis $\alpha_3 = 0$, and has asymptotically a χ^2 distribution with one degree of freedom (df) if $\alpha_3 = 0$. The LR test-of-fit for model 2 against model 4, a test of multiplicativity of effects, is given by

$$G^2(\beta_3 = 0) = -2[LL(2) - LL(4)]$$

$G^2(\beta_3 = 0)$ is a test statistic for the hypothesis $\beta_3 = 0$, and is asymptotically distributed as a χ^2 with one df if $\beta_3 = 0$.

Alternatively, Wald-type test statistics for additivity and multiplicativity obtain from the estimators of α_3 and β_3 and the estimated variances of these statistics. Specifically, suppose $\hat{\alpha}_3$ and

$\hat{\sigma}_\alpha^2$ are the maximum likelihood (ML) estimates of α_3 and the asymptotic variance of $\hat{\alpha}_3$; a test of additivity is then given by $Z^2(\alpha_3 = 0) = \hat{\alpha}_3^2/\hat{\sigma}_\alpha^2$, which is asymptotically distributed as a χ^2 with one df when $\alpha_3 = 0$. Similarly, a one df χ^2 test of multiplicativity is given by $Z^2(\beta_3 = 0) = \hat{\beta}_3^2/\hat{\sigma}_\beta^2$, where $\hat{\beta}_3$ and $\hat{\sigma}_\beta^2$ are the ML estimates of β_3 and the asymptotic variance of $\hat{\beta}_3$. $Z^2(\alpha_3 = 0)$ and $Z^2(\beta_3 = 0)$ are asymptotically equivalent to $G^2(\alpha_3 = 0)$ and $G^2(\beta_3 = 0)$, respectively; however, considerable differences can exist in their behaviour in small samples.[23] ('Asymptotically equivalent' is used here in the sense of Cox and Hinkley;[23] for the present discussion the important feature of such equivalence is that two equivalent tests may be regarded as having equal large-sample power.)

For the case of binary risk factors, $Z^2(\alpha_3 = 0)$ and $Z^2(\beta_3 = 0)$ are special cases of the omnibus χ^2 test of homogeneity described by Kupper and Hogan[6] and Fleiss,[22] and are also special cases of weighted least squares tests given by Grizzle, Starmer, and Koch.[21] In addition, $Z^2(\alpha_3 = 0)$ becomes identical to a test of additivity proposed by Hogan *et al.*[14] In fact, most of the methods proposed for testing homogeneity, 'effect modification', 'synergy' or 'interaction' reduce to one of the procedures described above when the risk factors are binary or continuous. One exception is a test of additivity based on Rothman's 'index of synergism';[2, 13] however, a simulation study by Hogan *et al.*[14] indicated that this test had poorer small-sample performance than the corresponding likelihood-ratio or Wald tests. Another important exception is a procedure given by Thomas[18] (described below) which is also useful for comparing the models with one another.

COMPARISONS OF ADDITIVITY AND MULTIPLICATIVITY

Having computed test statistics for the hypotheses of additivity ($\alpha_3 = 0$) and multiplicativity ($\beta_3 = 0$), it is natural to attempt to compare the two statistics [e.g. $G^2(\alpha_3 = 0)$ vs. $G^2(\beta_3 = 0)$ or $Z^2(\alpha_3 = 0)$ vs. $Z^2(\beta_3 = 0)$] to decide which of models 1 and 2 provides a better fit; one might choose the model with the less significant statistic. However, in the general case, models 3 and 4 are not equivalent; in other words, $G^2(\alpha_3 = 0)$ and $G^2(\beta_3 = 0)$ [and similarly $Z^2(\alpha_3 = 0)$ and $Z^2(\beta_3 = 0)$] do not test against the same alternative, and thus are *not* mutually comparable.[18]

To provide a unified framework for testing and comparing additive and multiplicative models, Thomas[18] proposes that the models be embedded in a 'mixture' model containing both models; one can then choose the model with the less significant test statistic, provided the fit of that model is acceptable. Cox and Hinkley describe a related method for choosing between models (Reference 23, p. 327).

Gardner and Munford[19] and Walker and Rothman[20] propose using a direct comparison of likelihoods in the course of choosing between models; for example, one may choose the model with the larger likelihood. This will lead to the same model choices as Thomas's procedures, for if two models are tested against a mixture, the model with the larger likelihood will have the less significant likelihood-ratio test statistic. Gardner and Munford and Walker and Rothman also propose that the relative fit of the models be directly measured by the magnitude of the likelihood ratio. In a comparison of models 1 and 2, this ratio is simply $\exp[LL(1) - LL(2)]$.

Given the above procedures for testing and comparing additive and multiplicative models, two questions arise: (i) what is the statistical power of the test statistics?; and (ii) what is the discrimination power of the model selection methods based on direct comparison of likelihoods? The specific answers to these questions depend heavily on the particulars of the situation, such as the strengths of the risk factors, joint distribution of the risk factors, scale of the risk factors (e.g. continuous, categorical, binary), baseline frequency of the disease, type of study design, and total sample size.

One special case of importance is that of a case-control study of two binary risk factors X and Y (both coded 0 = factor absent, 1 = factor present). A number of useful simplifications occur in this situation: Choosing $x_0 = 0$ and $y_0 = 0$ as the baseline categories, there are only three independent quantities to be estimated, namely $RR(1, 1)$, $RR(0, 1)$, and $RR(1, 0)$ ($RR(0, 0)$ being identically one). Assume that these may be validly estimated by the corresponding exposure-odds ratios from the case-control data. The tests for homogeneity of the log-odds ratio described by Fleiss,[22] Gart,[25] Berkson,[26] and Woolf[27] will then be identical to $Z^2(\beta_3 = 0)$, the likelihood-ratio test for interaction in a logistic or log-linear model will be identical to $G^2(\beta_3 = 0)$, and the Pearson χ^2 test for interaction in a log-linear model will be identical to the score test of $\beta_3 = 0$. Since models 3 and 4 each have three independent parameters, both models will be 'saturated' in this case and will perfectly fit the data, so that $LL(3) = LL(4)$ and $Z^2(\alpha_3 = 0) - Z^2(\beta_3 = 0) \simeq G^2(\alpha_3 = 0) - G^2(\beta_3 = 0) = -2[LL(1) - LL(2)]$ (where $\simeq$ denotes large-sample equivalence). Comparison of $Z^2(\alpha_3 = 0)$ and $Z^2(\beta_3 = 0)$ can thus serve to approximate likelihood comparison in this case.

POWER OF THE TESTS IN CASE-CONTROL STUDIES OF TWO BINARY RISK FACTORS

Let F be the cumulative distribution function of the standard normal distribution (zero mean, unit variance) and $c_{s/2}$ the $100(1-s/2)$ percentile of this distribution. Then a first-order asymptotic power function for the statistic $Z^2(\alpha_3 = 0)$ at significance level s is given by

$$\pi(s, \alpha_3, \sigma_\alpha^2) = F(-c_{s/2} - |\alpha_3|/\sigma_\alpha) + 1 - F(c_{s/2} - |\alpha_3|/\sigma_\alpha)$$

and an asymptotic power function for $Z^2(\beta_3 = 0)$ is given by

$$\pi(s, \beta_3, \sigma_\beta^2) = F(-c_{s/2} - |\beta_3|/\sigma_\beta) + 1 - F(c_{s/2} - |\beta_3|/\sigma_\beta)$$

To examine the performance of the tests for interaction, determine the useful range of the asymptotic power functions, and check the utility of the asymptotic 95 per cent confidence intervals $\hat{\alpha}_3 \pm 1{\cdot}96\hat{\sigma}_\alpha$ and $\hat{\beta}_3 \pm 1{\cdot}96\hat{\sigma}_\beta$, a number of simulation experiments were run. Table I presents a summary description of the ten models selected for presentation in this paper. They represent a series of situations in which one or both risk factors are of only moderate strength (odds ratio < 5) and so are situations in which tests of interaction might be expected to exhibit low power, even with moderately large samples. Such situations are common in epidemiologic research. Table II summarizes 30 simulation studies of 1000 trials each. In addition to $Z^2(\alpha_3 = 0)$ and $Z^2(\beta_3 = 0)$, study was made of Dayal's additivity test $Z_D^2(\alpha_3 = 0)$, a Wald test based on second-order approximations. Table II also presents simulation results of comparing $Z_D^2(\alpha_3 = 0)$ to $Z^2(\beta_3 = 0)$,

Table I. Description of models studied in Table II

		True odds ratios			True parameters	
Model	Type	$x = y = 1$	$x = 0, y = 1$	$x = 1, y = 0$	β_3	α_3
S1	Subadditive ($\alpha_3 < 0$)	6	6	3	$-1{\cdot}10$	-2
A1	Additive ($\alpha_3 = 0$)	6	4	3	$-0{\cdot}69$	0
A2	Additive	12	10	3	$-0{\cdot}91$	0
A3	Additive	18	16	3	$-0{\cdot}98$	0
I1	Intermediate ($\alpha_3 > 0$, $\beta_3 < 0$)	12	6	4	$-0{\cdot}69$	3
I2	Intermediate	18	10	3	$-0{\cdot}51$	6
M1	Multiplicative ($\beta_3 = 0$)	6	3	2	0	2
M2	Multiplicative	12	4	3	0	6
M3	Multiplicative	18	6	3	0	10
T1	Transmultiplicative ($\beta_3 > 0$)	12	3	3	$0{\cdot}29$	7

Table II

		Asymptotic and simulated powers*								
	Sample size	$Z^2(\beta_3 = 0)$		$Z(\alpha_3 = 0)$		$Z^2_D(\alpha_3 = 0)$		Coverage rates†		Proportion of
	(cases/	Asymp-	Simu-	Asymp-	Simu-	Asymp-	Simu-			trials with
Model	controls)	totic	lated	totic	lated	totic	lated	$\hat{\beta}_3 \pm 1{\cdot}96\hat{\sigma}_\beta$	$\hat{\alpha}_3 \pm 1{\cdot}96\hat{\sigma}_\alpha$	$Z^2(\beta_3 = 0) > Z^2_D(\alpha_3 = 0)$
S1	75/150	0·43 ,	0·42	0·13 ,	0·12	0·13 ,	0·13	0·93	0·94	0·12
	150/300	0·71 ,	0·70	0·21 ,	0·21	0·21 ,	0·22	0·95	0·94	0·04
	300/600	0·95 ,	0·94	0·36 ,	0·37	0·36 ,	0·37	0·95	0·93	0·01
A1	75/150	0·21 ,	0·22	0·05 ,	0·05	0·05 ,	0·06	0·94	0·95	0·30
	150/300	0·37 ,	0·36	0·05 ,	0·05	0·05 ,	0·05	0·96	0·95	0·21
	300/600	0·63 ,	0·64	0·05 ,	0·05	0·05 ,	0·05	0·94	0·95	0·14
A2	75/150	0·26 ,	0·25	0·05 ,	0·03	0·05 ,	0·03	0·97	0·97	0·24
	150/300	0·46 ,	0·47	0·05 ,	0·05	0·05 ,	0·05	0·95	0·95	0·15
	300/600	0·75 ,	0·76	0·05 ,	0·04	0·05 ,	0·04	0·96	0·96	0·08
A3	75/150	0·24 ,	0·20	0·05 ,	0·03	0·05 ,	0·04	0·96	0·97	0·27
	150/300	0·43 ,	0·46	0·05 ,	0·03	0·05 ,	0·04	0·97	0·97	0·14
	300/600	0·71 ,	0·71	0·05 ,	0·04	0·05 ,	0·04	0·96	0·96	0·08
I1	75/150	0·18 ,	0·18	0·13 ,	0·06	0·13 ,	0·08	0·96	0·96	0·42
	150/300	0·32 ,	0·30	0·21 ,	0·17	0·21 ,	0·19	0·95	0·96	0·42
	300/600	0.55 ,	0·58	0·37 ,	0·34	0·37 ,	0·35	0·94	0·95	0·38
I2	75/150	0·11 ,	0·07	0·20 ,	0·16	0·21 ,	0·20	0·96	0·93	0·59
	150/300	0·16 ,	0·17	0·36 ,	0·34	0·36 ,	0·35	0·95	0·93	0·63
	300/600	0·28 ,	0·29	0·62 ,	0·61	0·63 ,	0·62	0·96	0·93	0·67
M1	75/150	0·05 ,	0·05	0·18 ,	0·14	0·18 ,	0·16	0·95	0·96	0·69
	150/300	0·05 ,	0·05	0·32 ,	0·30	0·32 ,	0·32	0·95	0·96	0·79
	300/600	0·05 ,	0·05	0·55 ,	0·54	0·56 ,	0·55	0·95	0·95	0·86
M2	75/150	0·05 ,	0·06	0·44 ,	0·41	0·46 ,	0·45	0·94	0·93	0·84
	150/300	0·05 ,	0·05	0·72 ,	0·77	0·75 ,	0·78	0·96	0·92	0·94
	300/600	0·05 ,	0·06	0·95 ,	0·98	0·96 ,	0·98	0·95	0·91	0·98
M3	75/150	0·05 ,	0·04	0·55 ,	0·53	0·59 ,	0·57	0·96	0·88	0·88
	150/300	0·05 ,	0·05	0·84 ,	0·88	0·87 ,	0·89	0·95	0·86	0·96
	300/600	0·05 ,	0·04	0·986 ,	0·996	0·992 ,	0·996	0·96	0·87	0·994
T1	75/150	0·07 ,	0·06	0·59 ,	0·63	0·63 ,	0·67	0·96	0·90	0·94
	150/300	0·09 ,	0·09	0·87 ,	0·91	0·90 ,	0·91	0·95	0·90	0·98
	300/600	0·14 ,	0·13	0·992 ,	0·996	0.996 ,	0·996	0·95	0·89	0·998

* Simulated power will estimate true significance level for additivity tests under additive model (models A1, A2, A3) and multiplicativity test under multiplicative model (models M1, M2, M3); asymptotic power will equal nominal significance level under the null hypothesis.

† Percentage of trials interval covered true parameter (coverage rate for nominal 95 per cent confidence interval using ordinary maximum likelihood estimates of parameter and variance of estimator).

and the coverage rates of the 95 per cent confidence intervals for α_3 and β_3. Results are given here for sample sizes of 75 cases/150 controls ('small sample'), 150 cases/300 controls ('moderate sample'), and 300 cases/600 controls ('large sample'), using a nominal 0·05 significance level. The underlying joint risk-factor distribution of X and Yamong non-cases was fixed at $q_{11} = q_{01} = q_{10} = 0{\cdot}2$, $q_{00} = 0{\cdot}4$. Other sample sizes, case-control ratios, nominal significance levels, and risk factor distributions were studied; the simulations presented here illustrate the several important features of the statistics found in all instances.

The following points may be noted from Table II:

(1) In most instances, the asymptotic power function results and simulation results are in good agreement. Nearly all exceptions occur using the statistic $Z^2(\alpha_3 = 0)$ at the smallest sample size.
(2) The test $Z^2(\beta_3 = 0)$ is valid (i.e. rejects at very close to the nominal level) when a multiplicative model holds, even at the smallest sample size considered.
(3) The test $Z^2_{\mathrm{D}}(\alpha_3 = 0)$ (Dayal's test) is valid when an additive model holds, even at the smallest sample size considered.
(4) The power of $Z^2(\beta_3 = 0)$ is very low (below 0·50) when an additive or moderately trans-multiplicative model holds and the sample size is small or moderate.
(5) The powers of the additivity tests are low when a multiplicative or subadditive model holds and the sample size is small.
(6) All the tests have very low power when an intermediate model holds and the sample size is small or moderate; under such models the powers are not high even when using a large sample.
(7) $Z^2_{\mathrm{D}}(\alpha_3 = 0)$ shows somewhat better power and better agreement with its asymptotic power function than does $Z^2(\alpha_3 = 0)$.

The above results indicate that the statistics $Z^2(\beta_3 = 0)$ and $Z^2_{\mathrm{D}}(\alpha_3 = 0)$ are valid and that the asymptotic power functions given above provide good approximations to the true powers when the sample size is not very small (i.e. the minimum cell expectation is 4 or greater). This is consistent with Odoroff's[28] small-sample results regarding tests of multiplicativity. Nevertheless, the utility of a classical significance test is doubtful when the chance of rejecting a false null hypothesis under the alternative of primary interest is less than 50 per cent. In each of the additive situations presented, $Z^2(\beta_3 = 0)$ fails to reject multiplicativity in most of the trials involving the smallest or moderate sample sizes. The additivity tests appear to be somewhat more powerful in the analogous situations, but still exhibit low power in the trials involving an underlying multiplicative model and the smallest sample size. Thus, for a range of commonly encountered sample sizes and underlying models, failure to reject both the additive and multiplicative models will be a common occurrence. An asymptotic study of the tests using a cohort design led to the same conclusion.

The low power of the tests naturally leads to consideration of confidence intervals for the parameters as a means of expressing the imprecision in the sample information regarding interaction, and consideration of the likelihood comparison procedure for model selection. With regard to these approaches, the following points may be noted from Table II:

(1) The nominal 95 per cent confidence interval for β_3 exhibits coverage rates very close to the nominal rate (94–96 per cent coverage rates) under the models close to multiplicativity for even the smallest sample. Its performance under 'large' departures from multiplicativity ($|\beta_3| \geqslant 0{\cdot}7$) appears less satisfactory (93–97 per cent coverage rates) but is still adequate. ($|\beta_3| = 0{\cdot}7$ corresponds to a doubling of the odds ratio for X between strata of Y.)

(2) The 95 per cent confidence interval for α_3 performs adequately under models close to additivity (93–97 per cent coverage rates). Its performance deteriorates severely with increasing departures from additivity (coverage rates of 86–93 per cent when $|\alpha_3| \geqslant 5$).

(3) Comparison of $Z^2_D(\alpha_3 = 0)$ with $Z^2(\beta_3 = 0)$ appears to be a useful method for choosing between the additive and multiplicative models. For the smallest sample size, the procedure leads to selection of the correct model in most (69 per cent or more) of the trials. For the larger samples, the procedure is correspondingly better.

These results suggest that likelihood comparison provides a useful model selection criterion when the standard tests fail to reject either model. The accuracy of the model thus selected may then be judged with the aid of the confidence interval for the 'interaction' parameter (α_3 or β_3): the interval should, of course, contain zero; the more narrow and more centred about zero, the better is the accuracy. There remains, however, the technical problem of setting valid confidence limits for α_3, and further research into this area would be helpful.

DISCUSSION

It is worth noting that the tests of multiplicativity for several 2 × 2 tables (such as Woolf's test) can be validly applied if the data collection involved frequency-matching on Y (provided, of course, the smallest cell expectation is 'large'). This can be seen in the $2 \times 2 \times 2$ case by noting that the asymptotic expectation of $\hat{\beta}_3$ is not affected if the marginal distribution of Y (considered as a covariate) is fixed by design, as in matching. Unfortunately, the same property does not hold for tests of additivity: in general, matching on Y (or correlates of Y) will alter the asymptotic expectation of $\hat{\alpha}_3$ and lead to invalid tests of additivity.

Because of the asymptotic equivalence of the likelihood-ratio, score, and Wald tests, one would expect that the tests would exhibit practically the same power in large enough samples, and a simulation study of tests of multiplicativity in 3-way tables found this to be the case.[28] (Here, 'large enough sample' corresponds roughly to 'no cell expectation under four'.) From the closeness of the asymptotic and simulated powers in Table II it appears that the situations studied here are essentially asymptotic. Given these observations it is most likely that the true powers of the likelihood-ratio and score tests are close to the Wald tests powers in Table II, and thus the conclusion given earlier for the Wald tests regarding lack of power should apply to the likelihood-ratio and score tests as well. Nevertheless, it has been noted that under parameterizations such as those studied here the Wald statistic can exhibit aberrant behaviour, and thus the likelihood-ratio or score statistics are preferable in practice.[29-31]

With regard to small samples, earlier studies of likelihood-ratio tests of multiplicativity in 3-way tables indicate that these tests have poorer small-sample behaviour (in terms of rejection rates under the null hypothesis) than the corresponding score tests and Wald test,[28, 32] with the score test performing best. It appears from Table II, however, that in truly small samples the power of the tests will be so low as to render them useless.

Breslow and Day[33] have pointed out that the power of tests for interaction is much improved if the risk factors X and Y are continuous rather than binary. Another difficulty arises, however, if X and Y are continuous: it becomes necessary to specify the shape of the joint dose-response surface for X and Y. For example, it may not be known whether X and Y or some function of these (such as $\log X$ or $\log Y$) will be most appropriate for use as the regressors in the models discussed here. Specification of X and Y untransformed is the most common approach, but this is usually not justified on any grounds other than convenience. The dose-response question can be seen as adding new parameters to the problem in the continuous case, and these parameters must either be estimated or, as is more frequent in epidemiology, arbitrarily fixed. Thomas[18] addresses these issues in his discussion of model selection, and has provided some empirical examples of the problem.[34]

ACKNOWLEDGEMENTS

I received valuable advice from many colleagues in the course of preparation and revision of this paper; I would especially like to thank Robert Plourde, James J. Schlesselman, Duncan Thomas, and Alexander M. Walker for their detailed comments. I also thank Connie Brown and Virginia Hansen for manuscript preparation.

REFERENCES

1. Siegel, D. G. and Greenhouse, S. W. 'Multiple relative risk functions in case-control studies', *American Journal of Epidemiology*, **97**, 324–331 (1973).
2. Rothman, K. J. 'Synergy and antagonism in cause-effect relationships', *American Journal of Epidemiology*, **99**, 385–388 (1974).
3. Mantel, N., Brown, C., Byar, D. P. 'Tests for homogeneity of effect in epidemiologic investigations', *American Journal of Epidemiology*, **106**, 125–129 (1977).
4. Walter, S. D. and Holford, T. R. 'Additive, multiplicative, and other models', *American Journal of Epidemiology*, **108**, 341–346 (1978).
5. Rothman, K. J. 'Occam's razor pares the choice among statistical models', *American Journal of Epidemiology*, **108**, 347–349 (1978).
6. Kupper, L. L. and Hogan, M. D. 'Interaction in epidemiologic studies', *American Journal of Epidemiology*, **108**, 447–453 (1978).
7. Blot, W. J. and Day, N. E. 'Synergism and interaction—are they equivalent?', *American Journal of Epidemiology*, **110**, 99–100 (1979).
8. Hamilton, M. A. 'Choosing the parameter for a 2 × 2 table or a 2 × 2 × 2 table analysis', *American Journal of Epidemiology*, **109**, 362–375 (1979).
9. Siu, T. O. 'Letter to the Editor', *American Journal of Epidemiology*, **111**, 134–135 (1980).
10. Greenland, S. 'Limitations of the logistic analysis epidemiologic data', *American Journal of Epidemiology*, **110**, 693–698 (1979).
11. Saracci, R. 'Interaction and synergism', *American Journal of Epidemiology*, **112**, 465–466 (1980).
12. Rothman, K. J., Greenland, S. and Walker, A. M. 'Concepts of Interaction', *American Journal of Epidemiology*, **112**, 467–470 (1980).
13. Rothman, K. J. 'The estimation of synergy or antagonism', *American Journal of Epidemiology*, **103**, 506–511 (1976).
14. Hogan, M. D., Kupper, L. L., Most, B. M. and Haseman, J. K. 'Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies', *American Journal of Epidemiology*, **108**, 60–67 (1978).
15. Rothman, K. J. 'Estimation versus detection in the assessment of synergy', *American Journal of Epidemiology*, **108**, 9–11 (1978).
16. Hogan, M. D., Haseman, J. K., Kupper, L. L. and Most, B. M. 'Letter to the Editor', *American Journal of Epidemiology*, **108**, 159–160 (1978).
17. Dayal, H. H. 'Additive excess risk model for epidemiologic interaction in retrospective studies', *Journal of Chronic Diseases*, **33**, 653–660 (1980).
18. Thomas, D. C. 'General relative risk models for survival time and matched case-control studies', *Biometrics*, **37**, 673–686 (1981).
19. Gardner, M. J. and Munford, A. G. 'The combined effect of two factors on disease in a case-control study', *Applied Statistics*, **29**, 276–281 (1980).
20. Walker, A. M. and Rothman, K. J. 'Models of varying parametric form in case-referent studies', *American Journal of Epidemiology*, **115**, 129–137 (1982).
21. Grizzle, J. E., Starmer, C. F. and Koch, C. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489–504 (1969).
22. Fleiss, J. L. *Statistical Methods for Rates and Proportions* (2nd edition), Wiley, New York, 1981.
23. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974.

24. Wald, A. 'Tests of statistical hypotheses concerning several parameters when the number of observations is large', *Transactions of the American Mathematical Society*, **54**, 426–482 (1943).

25. Gart, J. J. 'The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification', *International Statistical Institute Review*, **39**, 148–169 (1971).

26. Berkson, J. 'Application of minimum-logit χ^2 estimate to a problem of Grizzle', *Biometrics*, **24**, 75–95 (1968).

27. Woolf, B. 'On estimating the relation between blood group and disease', *Annals of Human Genetics*, **19**, 251–253 (1954).

28. Odoroff, C. L. 'A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2 × 2 × 2 and 3 × 2 × 2 contingency tables: tests for interaction', *Journal of the American Statistical Association*, **65**, 1617–1631 (1970).

29. Hauck, W. W. and Donner, A. 'Wald's test as applied to hypotheses in logit analysis', *Journal of the American Statistical Association*, **72**, 851–853 (1977).

30. Hauck, W. W. and Donner, A. 'Corrigenda', *Journal of the American Statistical Association*, **75**, 482 (1980).

31. Vaeth, M. 'On the use of Wald's test in exponential families', *Research Report no. 70*, Dept. of Theoretical Statistics, University of Aarhus, Denmark, 1981.

32. Larntz, K. 'Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics', *Journal of the American Statistical Association*, **73**, 253–263 (1978).

33. Breslow, N. E. and Day, N. E. 'Analysis of case-control studies', *IARC Publication no. 32*, WHO Publications, Geneva, 1981.

34. Thomas, D. C. 'Are dose-response, synergy and latency confounded?', *Presentation to the Joint Statistical Meetings of ASA and the Biometric Society*, Detroit, July 1980.