Statistics in Epidemiology: The Case-Control Study

N. E. Breslow

Statistics in Epidemiology: The Case-Control Study

N. E. BRESLOW

Statisticians have contributed enormously to the conceptualization, development, and success of case-control methods for the study of disease causation and prevention. This article reviews the major developments. It starts with Cornfield's demonstration of odds ratio invariance under cohort versus case-control sampling, proceeds through the still-popular Mantel–Haenszel procedure and its extensions for dependent data, and highlights (conditional) likelihood methods for relative risk regression. Recent work on nested case-control, case-cohort, and two-stage case-control designs demonstrates the continuing impact of statistical thinking on epidemiology. The influence of R. A. Fisher's work on these developments is mentioned wherever possible. His objections to the drawing of causal conclusions from observational data on cigarette smoking and lung cancer are used to introduce the problems of measurement error and confounding bias. The resolution of such difficulties, whether by further development and implementation of randomized intervention trials or by causal analysis of observational data using graphical models containing latent variables, will challenge future generations of statisticians.

KEY WORDS: Likelihood; Mantel–Haenszel procedure; Matched samples; Observational data; Odds ratio; Relative risk regression.

1. INTRODUCTION

The sophisticated use and understanding of case-control studies is the most outstanding methodologic development of modern epidemiology (Rothman 1986, p. 62).

My choice of topic for the 1995 Fisher Lecture is based on my belief that the contributions made by statisticians to the development of case-control methodology over the past 50 years have been among the most important of the many contributions they have made to public health and biomedicine. This view is shared by many epidemiologists. Writing in the first 1994 issue of *Epidemiologic Reviews*, which was devoted entirely to applications of the case-control method, Armenian and Lilienfeld (1994, p. 3) declared that the impact of statisticians on the "development of epidemiology would be difficult to overstate." Rothman's quotation, from his influential textbook *Modern Epidemiology*, highlights the importance of case-control methods in current epidemiologic research. The continuing popularity of the methodology is evident from the fact that 223 population-based case-control studies were published in the world literature in 1992 (Correa, Stewart, Yeh, and Santos-Burgoa 1994).

I am most grateful to the Committee and to the Organizers for the invitation to present the 1995 Fisher Lecture and for the opportunity to discuss a subject that has stimulated much of my research work. I would like to acknowledge Professors L. Moses and B. Efron, my graduate and dissertation advisors; Professor P. Armitage, who hosted me during a seminal postdoctoral year; and above all Professor N. Day, who introduced me to case-control studies and with whom I have enjoyed a long and fruitful collaboration. It is also a pleasure to acknowledge the outstanding contributions made to this field, and to my understanding of it, by my colleagues and by a score of graduates of the University of Washington Biostatistics Program.

2. ORIGINS

The central idea of the case-control study is the comparison of a group having the outcome of interest to a control group with regard to one or more characteristics. An early example is Guy's 1843 comparison of the occupations of men with pulmonary consumption to those of men with other diseases (Lilienfeld and Lilienfeld 1979). The method became popular during the 1920s for the study of cancer, notable successes being the associations discovered between lip cancer and pipe smoking by Broders (1920), between breast cancer and reproductive history by Lane-Claypon (1926), and between oral cancer and pipe smoking by Lombard and Doering (1928). Because these diseases were rare, it was rather impractical to study them in any other way; for example, by follow-up of an initially healthy population. Increased attention to *and criticism of* case-control methodology followed the publication in 1950 of several studies of smoking and lung cancer (Surgeon General 1964).

Under the leadership of Harold Dorn, statisticians at the U.S. National Cancer Institute were stimulated by the ensuing controversy to investigate the advantages and shortcomings of the case-control method. A prevailing belief at the time was that separate samples of cases and controls did not provide relevant quantitative information about the parameters of primary interest—namely, the disease rates. This misconception was corrected by Jerome Cornfield (1951), who is widely credited with launching the modern era of case-control studies. Cornfield demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed, and that the latter in turn approximates the ratio of disease rates provided that the disease is rare. Formally, if D denotes disease (1 for cases, 0 for controls) and X denotes exposure (1 for

exposed, 0 unexposed), then

$$\frac{\Pr(X = 1|D = 1)\Pr(X = 0|D = 0)}{\Pr(X = 1|D = 0)\Pr(X = 0|D = 1)}$$

$$= \frac{\Pr(D = 1|X = 1)\Pr(D = 0|X = 0)}{\Pr(D = 1|X = 0)\Pr(D = 0|X = 1)}$$

$$\approx \frac{\Pr(D = 1|X = 1)}{\Pr(D = 1|X = 0)}. \tag{1}$$

The exposure odds ratio, now widely known as the relative risk, is thus understood to approximate the disease rate ratio. Cornfield further demonstrated that if the disease rate in the general population is known, then it may be combined with the case-control data to yield separate estimates of disease rates for exposed and unexposed. He was well aware of the need for the case and control groups to be "representative of these same groups in the general population" (Cornfield 1951, p. 1273) for his calculations to be valid.

The disease rates to which Cornfield referred were *prevalence* rates—simple proportions of individuals having the disease, as would be obtained in a cross-sectional sample. For studies of disease etiology, however, it is preferable to study *incidence* rates and to estimate the (ratios of) probabilities of disease development during a specified time period for individuals who are disease free at its start. Otherwise, one runs the risk of confusing the effects of exposure on the disease incidence rate with its effects on the case fatality rate (Neyman 1955). We now know that even without the rare disease assumption, the exposure odds ratio that is estimable from a case-control study approximates the ratio of instantaneous disease incidence rates, provided that the controls are sampled proportionately with the incident cases throughout the study period (see Sec. 5).

Cornfield's demonstration did not quiet all the critics, one of the most vociferous being R. A. Fisher (1957a,b; 1958a,b). Fisher raised the issue of association versus causation that clouds the interpretation of any observational study. In his famous constitutional hypothesis, he suggested that the smoking and lung cancer association could be explained by the confounding effects of a genotype that predisposed both to smoking and to lung cancer. Data on twins were used to substantiate his assertion that smoking behavior was influenced by genetics. Cornfield and colleagues (1959) responded to these charges and others in a lengthy review that is well worth reading for its insights regarding causal inference. A centerpiece of their argument was a simple calculation showing that for a confounding factor to explain a relative risk of a given magnitude, say ψ, this factor had to be ψ times more prevalent among the exposed than among the unexposed. Because the lung cancer relative risk was approximately 10 for cigarette smokers versus nonsmokers and increased to 20 for heavy smokers, the existence of such a confounding factor seemed quite implausible. This calculation was later formalized into the concept of the confounding risk ratio, which measures the possible extent of confounding on the observed relative risk

(Miettinen 1972; Schlesselman 1978). These developments further solidified relative risk as a meaningful parameter for epidemiologic study.

By the end of the decade, the desiderata for case-control studies that could yield credible results with implications for possible public health action were fairly well understood. The main points, as summarized by Dorn in his 1959 Cutter Lecture at Harvard (Dorn 1959), were as follows. The study should (a) be conducted in a defined population, (b) include all incident cases occurring during a specified time period, (c) utilize objective measures of exposure to putative risk factors; (d) use multiple control groups, (e) be replicated in different populations, and (f) be verified by a cohort study. This last point suggests that Dorn was still not entirely convinced by the theoretical arguments of his colleagues that the case-control approach was as valid as the prospective approach that he had pioneered. I return to his concerns, and to those of Fisher, in the concluding section.

3. THE MANTEL–HAENSZEL ERA

Epidemiologists who have done case-control studies during the past 20 years ... have stood on the shoulders of giants. And, lest we epidemiologists lose sight of one major root of our discipline, we should remember that all of these men are, or were, statisticians (Cole 1979, p. 15).

The statisticians to whom Cole refers are Cornfield and Dorn and their colleagues Mantel and Haenszel, who in 1959 published their landmark paper in the *Journal of the National Cancer Institute*. This paper clarified the relationship between case-control (or retrospective) and cohort (forward or prospective) studies with the observation that "*a primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done*" (Mantel and Haenszel 1959, p. 733). Anticipating the development of the nested case-control study (see Sec. 5), Mantel and Haenszel suggested that one might adopt the case-control approach even to the sampling of subjects already ascertained in a cohort study, to collect additional data items. Clearly, the only *conceptual* difference between cohort and case-control studies was that the latter involved sampling from the cohort rather than complete enumeration of it.

Two statistical procedures were introduced for the control of confounding by stratification of the data into a series of 2×2 tables. The first was an adjusted chi-squared test that elaborated on earlier work by Cochran (1954). The second was a summary relative risk estimator intended to weight the individual odds ratios by both precision and importance. Referring to Table 1 for basic notation, the Mantel–Haenszel (MH) summary relative risk estimator is defined by the simple formula $\hat{\psi}_{\mathrm{MH}} = \sum_i R_i / \sum_i S_i$, where $R_i = A_i D_i / N_i$ and $S_i = B_i C_i / N_i$. The recommended procedures were eventually adopted for routine use by epidemiologists, who benefited from seeing their data arranged in tabular form and from making comparisons of individual and summary relative risks that alerted them to possible heterogeneity. The paper had enormous impact. By the end of 1994, it had received more than 4,000 citations, being one of the 200 most cited papers in the scientific literature

Table 1. Notation for a Series of 2 × 2 Tables

Exposure	Observed frequencies			Expected values	
	Case $D = 1$	Cont $D = 0$	Total	Case $D = 1$	Cont $D = 0$
$X = 1$	A_i	B_i	t_i	$N_{1i}p_{1i}$	$N_{0i}p_{0i}$
$X = 0$	C_i	D_i	$N_i - t_i$	$N_{1i}q_{1i}$	$N_{0i}q_{0i}$
Total	N_{1i}	N_{0i}	N_i	N_{1i}	N_{0i}

Odds ratio: $\psi_i = (p_{1i}q_{0i})/(q_{1i}p_{0i})$

since 1945. It continues to receive citations at the rate of about 250 per year, more today than during the 1970s and 1980s (Institute for Scientific Information, personal communication, Philadelphia).

Mantel and Haenszel presented no variance formula for their estimator and referred to work by Cornfield (1956) for calculation of interval estimates. It took 25 years to develop a simple, robust formula. Part of the problem was the fact that the estimator was useful, and was being used, in two rather different asymptotic environments: I, a small number of tables with large frequencies, and II, a large number of tables with small frequencies, such as arise from finely stratified or matched studies. Building on earlier work by Hauck (1979) and Breslow (1981) for asymptotics I and II a variance estimate that encompassed these two situations and intermediate ones was developed independently by Robins, Breslow, and Greenland (1986) and by Phillips and Holland (1987). The key observations are as follows. First, $E(R_i) = \psi_i E(S_i)$, where ψ_i denotes the true odds ratio in table i. Assuming a common value for the ψ_i, and denoting $R = \sum_i R_i$ and $S = \sum_i S_i$, $\hat{\psi}_{\text{MH}}$ is thus the solution of the unbiased estimating equation $R - \psi S = 0$. Second, under paired binomial sampling, the variances of the individual contributions to this estimating equation satisfy

$$N_i^2 \text{var}(R_i - \psi S_i)$$
$$= \frac{1}{2} E\{(A_i D_i + \psi B_i C_i)[A_i + D_i + \psi(B_i + C_i)]\}. \quad (2)$$

Finally, with $\beta = \log(\psi)$ denoting the log relative risk,

$$\hat{\beta}_{\text{MH}} = \log \hat{\psi}_{\text{MH}} = \beta + \frac{R - \psi S}{E(R)} + o_p\left(\frac{\text{var}(R)}{E^2(R)} + \frac{\text{var}(S)}{E^2(S)}\right).$$

Table 2. Regression Coefficients and Standard Errors

Variable	CML*	M-H procedure
Constant odds ratio		
Intercept	.924 ± .132	.903 ± .152
Odds ratio regression model		
Intercept	4.440 ± 5.009	4.896 ± 4.967
Female sex	−.591 ± 1.001	−.764 ± .976
White race	−2.625 ± 1.056	−2.595 ± 1.015
Age (yr)	−.021 ± .079	−.028 ± .077
Hospital status	1.060 ± .888	1.221 ± .758

* Conditional maximum likelihood.
Source: Liang, Beaty, and Cohen (1986). Reprinted with permission.

Combining the last two equations yields the variance estimator

$$\hat{\text{var}}(\hat{\beta}_{\text{MH}}) = \frac{1}{R^2} \sum_i N_i^{-2}(A_i D_i + \hat{\psi}_{\text{MH}} B_i C_i)$$
$$\times [A_i + D_i + \hat{\psi}_{\text{MH}}(B_i + C_i)].$$

Alternative computing formulas have been given by Robins, Greenland, and Breslow (1986).

When homogeneity of the odds ratio is in doubt, it is sensible to model the relative risk as a function of covariables associated with each stratum; for example, $\log \psi_i = z_i \beta$. Davis (1985) and Liang (1985) each suggested estimating equations for the regression coefficients β of the form

$$\sum_i w_i(\psi_i)(R_i - \psi_i S_i)z_i = 0, \quad (3)$$

where the w_i are appropriately selected weights ($w_i = \psi_i^{-1/2}$ for Liang). Cologne and Breslow (1990) recommended basing the variance of the resulting estimator on (2) and demonstrated through simulations that this extended MH procedure maintained good efficiency relative to maximum likelihood when the binomial sampling model held.

A major advantage of the MH procedure is that the equations (3) are unbiased and the corresponding estimate is consistent, even when the observations within strata are *dependent* and the binomial sampling model fails (Liang 1985, 1987). In that case, one uses a "sandwich"-type empirical variance estimator (Huber 1967). The approach is ideal for case-control studies of familial aggregation where A_i and B_i represent the number of family members of the case and of the control who are affected by the same disease as the case. Liang, Beaty, and Cohen (1986) gave an example involving chronic obstructive pulmonary disease. They tested 79 first-degree relatives of 28 cases and found 33 who had impaired pulmonary function. The number affected was only 15 among 77 relatives of 28 controls matched on sex, age, and hospital status (inpatient vs. outpatient). Family members of cases had a 2.5-fold greater disease prevalence overall than those of controls, and there was some evidence for a stronger association among blacks (see Table 2). Note the increased variance when the possible dependence is taken into account by the extended MH procedure. Donner and Hauck (1988) used an estimate of the intrafamilial correlation in a related procedure designed to augment efficiency.

Some statisticians may have concluded that the Mantel–Haenszel era ended with the introduction of logistic regression. They would be wrong. As Kahn and Sempos rightly remarked in their 1989 textbook *Statistics in Epidemiology*, "when a method is as simple and free of assumptions as the M–H procedure, it deserves to receive a strong recommendation, and we do not hesitate to give it" (Kahn and Sempos 1989, p. 156).

4. LIKELIHOODS AND LOGITS

4.1 Inference on Odds Ratios

The aspect of R. A. Fisher's work that has had the greatest impact on modern approaches to the analysis of case-control data is his development of likelihood inference based on explicit probability models (Fisher 1922). The centrality of likelihood is emphasized in the 1993 text *Statistical Models in Epidemiology* (Clayton and Hills 1993), in which chapter headings include "Likelihood," "Approximate Likelihoods," "Likelihoods for the Rate Ratio," and "Likelihoods for the Odds Ratio." Fisher (1935) himself introduced likelihood inference for the odds ratio in a 2×2 table in his classic paper on the logic of inductive inference. There he considered rates of criminality among same-sex twins of known criminals according to whether the twins were monozygotic or dizygotic. The key feature of his analysis was the use of the conditional distribution of the paired binomial data given the marginal totals, which he considered to be *ancillary* in the sense that they supplied no information about the parameter of interest but did indicate how precisely that parameter could be estimated. The resulting distribution, whose only parameter is the unknown odds ratio ψ, is known today as the extended hypergeometric distribution (Harkness 1965). For the setup in Table 1,

$$\Pr(A_i = a_i | t_i; \psi_i) \propto \binom{N_{1i}}{a_i} \binom{N_{0i}}{t_i - a_i} \psi_i^{a_i}. \quad (4)$$

Cornfield (1956) approximated exact confidence limits for ψ_i based on (4) by taking A_i to be normally distributed with mean $\hat{A}_i$, the fitted value determined by the marginal totals and the unknown ψ_i as the solution to the quadratic equation

$$\frac{\hat{A}_i (N_{0i} - t_i + \hat{A}_i)}{(N_{1i} - \hat{A}_i)(t_i - \hat{A}_i)} = \psi_i \quad (5)$$

and asymptotic variance

$$\text{var}_i(\hat{A}_i)$$
$$= \left(\frac{1}{\hat{A}_i} + \frac{1}{N_{1i} - \hat{A}_i} + \frac{1}{t_i - \hat{A}_i} + \frac{1}{N_{0i} - t_i + \hat{A}_i} \right)^{-1}. \quad (6)$$

The procedure was extended to interval estimation of the common odds ratio from a series of tables by Gart (1970). Exact confidence limits for the common value, which involve the convolution of the distributions (4), are now available in commercial software that implements the network algorithm of Mehta, Patel, and Gray (1985).

My own work with Fisher's distribution has involved relative risk regression models of the form $\log \psi_i = z_i \beta$ (Breslow 1976). The conditional maximum likelihood (CML) estimate of β is obtained from the conditional likelihood equations

$$\sum_i \{A_i - E[A_i | t_i; \psi_i = e^{z_i \beta}]\} z_i = 0.$$

A favorite example is a case-control study of childhood cancer and in utero irradiation conducted in the Oxford region (Stewart and Kneale 1970). Data for more than 6,000 children who had died from cancer during 1956–1964 at ages 0–9 years, and for an equal number of controls, were arranged in a series of 120 2×2 tables stratified by year of birth and the age at death of the case. Exposure was positive if the mother reported that she had received pelvic irradiation during pregnancy. The regression analysis showed clearly that age had no effect on the relative risk, thus contradicting a claim that the "radiogenic" cases had a more peaked age distribution than the "idiopathic" cases (Kneale 1971). But there was clear evidence for a decline in relative risk with calendar year of birth, which was consistent with the declining dose levels of medical irradiation. What seemed less clear was whether the decline might have attenuated with time. Although the addition of a quadratic term in birthyear significantly increased the likelihood, I wondered whether possible overdispersion, in the form of excess scatter of the individual relative risk estimates about the regression line, had been adequately accounted for.

An opportunity to investigate this question presented itself during recent work with Clayton on approximate inference procedures for generalized linear mixed models (Breslow and Clayton 1993). We included in the linear predictor a random birth year effect, assumed to have a normal distribution with mean zero and variance σ^2. The evidence for overdispersion was equivocal, with $\hat{\sigma} = .15 \pm .10$. Inclusion of the independent random error terms in the model had little effect on the statistical significance of the fixed, quadratic term, thus confirming the original conclusion regarding the inadequacy of a simple linear model in birth year. Further analysis using essentially the same model, but with an autoregressive correlation structure specified for the random effects to stochastically smooth the regression curve, showed a flattening of the relative risk during the mid 1950s. This seemed considerably more plausible than the upturn predicted by the parametric model (Fig. 1).

4.2 The "Breslow–Day" Test of Odds Ratio Homogeneity

The conditional likelihood score test for homogeneity of the odds ratio against global alternatives of heterogeneity takes the form

$$\sum_i \frac{[A_i - E(A_i | t_i; \hat{\psi}_{\text{CML}})]^2}{\text{var}(A_i | t_i; \hat{\psi}_{\text{CML}})}. \quad (7)$$
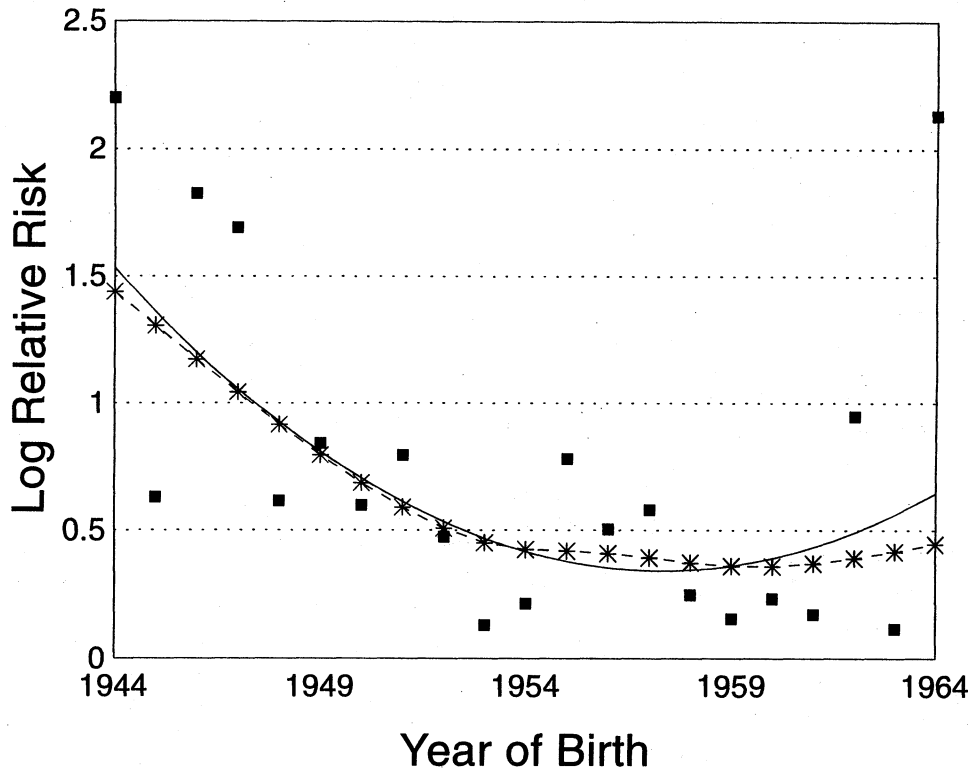
Figure 1. Relative Risks of Childhood Cancer in the Oxford Region for Children Exposed to In Utero Irradiation Versus the Nonexposed, by Year of Birth. The logarithms of the fitted relative risks are plotted for the fixed effects model as filled squares (■) and for the autoregressive random effects model as asterisks (∗). The curved line (——) represents the fitted relative risks in a fixed effects model with linear and quadratic terms for year of birth. (Source: Breslow and Clayton 1993, Fig. 4.)

Under asymptotic model I, an asymptotically equivalent statistic replaces the exact conditional means and variances with the approximations (5, 6) and the CML estimate with the "asymptotic" maximum likelihood (AML) estimate based on the unconditional likelihood. This asymptotic version is simply the standard chi-squared goodness-of-fit test in disguise. In my IARC monograph with Day on case-control studies (Breslow and Day 1980), the rather casual *and incorrect* assertion was made that one could substitute $\hat{\psi}_{\mathrm{MH}}$ for the iterative AML estimate. When using an inefficient estimator in such statistics, it is necessary to adjust the scores for the parameters to be tested, here representing heterogeneity in the odds ratios, by subtracting off their regression on the score(s) for the remaining parameter(s), here the common odds ratio (Neyman 1959). Using these principles, the correct form for the heterogeneity test was derived by Tarone (1985) as

$$\sum_i \frac{(A_i - \hat{A}_{i,\mathrm{MH}})^2}{\mathrm{var}_i(\hat{A}_{i\mathrm{MH}})} - \frac{(\sum_i A_i - \sum_i \hat{A}_{i,\mathrm{MH}})^2}{\sum_i \mathrm{var}_i(\hat{A}_{i,\mathrm{MH}})} , \qquad (8)$$

where $\hat{A}_{i,\mathrm{MH}}$ denotes the fitted frequencies estimated from the quadratic equation (5) with $\psi_i = \hat{\psi}_{\mathrm{MH}}$ and $\mathrm{var}_i(\hat{A}_{i,\mathrm{MH}})$ is the variance obtained by substituting these frequencies into (6). Fortunately, the MH estimator is so nearly efficient that the correction term in (8) is frequently negligible. Simulation studies report comparable properties for the uncorrected and corrected versions of the test (Jones, O'Gorman, Lemke, and Woolson 1989). Rather unfortunately, the mis-

take has made its way into the computer packages, and one sees increasing reference in the epidemiology literature to the asymptotically invalid "Breslow–Day" test for homogeneity of the odds ratio (Cytel Software Corporation 1995). I hope that the matter will be rectified soon and that Tarone's statistic becomes the standard of the future. Liang and Self (1985) developed a homogeneity test for the sparse data (asymptotics II) situation.

4.3 Logistic Regression

Log odds ratio regression models concern the effects of a single binary risk factor on disease risk. Multiple categorical risk factors may be accommodated, but only by considering each factor after stratification to control for effects of the others, and only by considering each level of exposure separately relative to baseline. Methods to evaluate the *simultaneous* effects of multiple quantitative risk factors on disease rates began to be developed during the 1960s, stimulated by the requirements of several large cohort studies of cardiovascular disease, particularly the Framingham study. Once again it was Cornfield who led the way with an application of Fisher's (1936) linear discriminant to the analysis of the Framingham data (Cornfield, Gordon, and Smith 1961). The goal was not simply to discriminate between two populations; rather, it was to summarize, in a simple mathematical form, the risk of developing disease during a specified time period as a function of one or more exposure variables measured for each person at the start. Cornfield noted that if the multivariate distributions of exposure X

among persons with and without disease were normal, with separate means but a common covariance matrix, then the probability of developing disease for an individual with values $\mathbf{X} = \mathbf{x}$ was given by the logistic response curve

$$\Pr(D = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\alpha + \mathbf{x}\beta)}{1 + \exp(\alpha + \mathbf{x}\beta)} . \qquad (9)$$

The parameters (α, β) are simple functions of the moments of the normal exposure distributions and of the marginal (prior) probability of disease development. Cornfield proposed to estimate them by the corresponding sample quantities; that is, by the linear discriminant.

Cox (1966) recommended instead that one estimate the parameters by maximum likelihood using only the logistic specification (9), which involves fewer assumptions. By allowing the exposure distribution in the control population to be completely arbitrary, Day and Kerridge (1967) noted more formally that the full likelihood based on the joint distribution of $(D, \mathbf{X})$ could be factored into the product of (a) the conditional likelihood, specified by the logistic model, and (b) the marginal likelihood of the exposures, and that both pieces could be maximized separately. Thus they confirmed that logistic regression, as we know it today, was efficient in the modern, semiparametric sense.

A key feature of the logistic model for case-control studies is that the regression coefficients β have a relative risk interpretation (Seigel and Greenhouse 1973). Formally,

$$\frac{\Pr(D = 1|\mathbf{X} = \mathbf{x}_1)\Pr(D = 0|\mathbf{X} = \mathbf{x}_0)}{\Pr(D = 0|\mathbf{X} = \mathbf{x}_1)\Pr(D = 1|\mathbf{X} = \mathbf{x}_0)} = \exp[(\mathbf{x}_1 - \mathbf{x}_0)\beta],$$

so that $(\mathbf{x}_1 - \mathbf{x}_0)\beta$ represents the log relative risk for a subject with exposures $\mathbf{x}_1$ versus one with exposures $\mathbf{x}_0$. The only complication is that the likelihood for case-control sampling contains terms of the form $\Pr(\mathbf{X}|D)$ rather than $\Pr(D|\mathbf{X})$. Anderson (1972) noted the difficulty and solved the problem for a discrete exposure distribution taking K values $\mathbf{x}_1, \ldots, \mathbf{x}_K$, as follows. Suppose that one samples n_1 cases and n_0 controls for a total sample size of $n = n_0 + n_1$, and observes n_{1k} cases and n_{0k} controls with $\mathbf{X} = \mathbf{x}_k$. Denote the probabilities specified by the logistic model (9) by $p_{1k}^* = 1 - p_{0k}^* = \Pr(D = 1|\mathbf{X} = \mathbf{x}_k)$, and the marginal probabilities of exposure by $q_k^* = \Pr(\mathbf{X} = \mathbf{x}_k)$. Using the fact that $\Pr(\mathbf{X}|D) = \Pr(D|\mathbf{X})\Pr(\mathbf{X})/\Pr(D)$, and assuming that the marginal disease probabilities $\Pr(D = i) = \pi_i$ are known, the case-control likelihood is proportional to

$$L_1^* L_2^* = \left\{ \prod_{i=0}^{1} \prod_{k=1}^{K} (p_{ik}^*)^{n_{ik}} \right\} \left\{ \prod_{k=1}^{K} (q_k^*)^{n_{+k}} \right\}.$$

But instead of involving free parameters as in the Day–Kerridge formulation, the parameters are *constrained* by the fixed marginal probabilities of disease: $\sum_k p_{ik}^* q_k^* = \pi_i$ for $i = 0, 1$.

Anderson solved the constrained estimation problem using the classical theory of Aitchison and Silvey (1958). He discovered that the estimates and covariance matrix for the relative risk coefficients β were identical to those of ordinary logistic regression involving maximization of L_1^* alone. He also discovered that the algebra was eased if one could

assume that the marginal disease probabilities π_i were equal to the relative frequencies n_i/n of cases and controls, and furthermore that the assignment of any specific values to the π_i left the relative risk estimates and their variances unchanged.

Anderson's approach via constrained maximum likelihood estimation was not strictly valid for continuous exposure variables, because the number of nuisance parameters, the q_k^*, increased with the sample size. Prentice and Pyke (1979) constructed a proof of his results that applied more generally. Their starting point was another factorization of the likelihood

$$L_1 L_2 = \left\{ \prod_{i=0}^{1} \prod_{k=1}^{K} (p_{ik})^{n_{ik}} \right\} \left\{ \prod_{k=1}^{K} (q_k)^{n_{+k}} \right\},$$

where now

$$q_k = \sum_i \frac{n_i}{n} \Pr(\mathbf{X} = \mathbf{x}_k|D = i)$$

represents the probability that a randomly chosen member of the case-control sample has exposures $\mathbf{x}_k$ and likewise, with $\gamma = \alpha + \log[(n_1 \pi_0)/(n_0 \pi_1)]$,

$$p_{1k} = 1 - p_{0k} = \frac{p_{1k}^* q_k^* n_1/\pi_1}{\sum_i p_{ik}^* q_k^* n_i/\pi_i} = \frac{\exp(\gamma + \mathbf{x}_k \beta)}{1 + \exp(\gamma + \mathbf{x}_k \beta)}$$

is the probability that a sample member with $\mathbf{X} = \mathbf{x}_k$ is a case. The parameters are now constrained by the requirements that the marginal probabilities of being a case or a control are fixed by design: $\sum_k p_{ik} q_k = n_i/n$ for $i = 0, 1$.

Prentice and Pyke (1979) demonstrated that the solution to the *unconstrained* maximization problem, with $(\hat{\gamma}, \hat{\beta})$ the ordinary logistic regression coefficients based on L_1 and $\hat{q}_k = n_{+k}/n$ the sample $\mathbf{X}$ distribution, actually satisfied the constraints and thus yielded the desired estimates. They further showed that the estimating equations derived from L_1 were unbiased and, using estimating equation theory, confirmed that the usual covariance matrix for $\hat{\beta}$ remained valid under case-control sampling. Because the intercept γ was a free parameter, it did not matter that the π_i's were unknown. Carroll, Wang, and Wang (1995) recently extended the basic Prentice–Pyke results regarding validity of fitting "prospective" logistic regression models to case-control data to related procedures for robust estimation of regression coefficients, correction of measurement error, and partially missing data.

While this work was in progress, a parallel and in some respects deeper investigation of methods for binary response models with outcome dependent sampling was underway in econometrics (Cosslett 1981; Manski and Lerman 1977). Here one refers to "choice-based" rather than "case-control" samples, because the outcomes involve economic choices rather than cases of disease. Auxiliary data on the marginal choice probabilities π_i were generally assumed to be available, which facilitated the study of quantal response models other than the logistic. An important point emphasized in the social science literature, and somewhat

neglected by biostatisticians and epidemiologists, is the sensitivity of maximum likelihood estimates to misspecification of the population model. If one posits a linear logistic model for the disease rates in the population but the true model is quadratic, for example, then the regression coefficient estimated from the case-control sample may differ substantially from the coefficient that one would estimate from a cohort study of the same population (Scott and Wild 1986; Xie and Manski 1989). When the marginal disease rates in the population are known, it often is preferable to use an inefficient but more robust estimate that *weights* the usual likelihood contributions by the inverse sampling fractions, which of course differ greatly between cases and controls (Manski and McFadden 1981). Table 3 shows properties of the weighted and unweighted estimators in a linear logistic model where the exposure variable has a standard normal distribution and where the true model for the log relative risk is the quadratic $-5.1 + 2.0x + .3x^2$. Due to the curvature, the optimal β when fitting the linear equation to the population is 2.8 rather than 2.0. Relative to this value, the weighted likelihood estimator clearly outperforms "maximum likelihood."

5. MATCHING AND NESTING

I was introduced to the case-control design in 1972 during collaborative work at IARC on a study of esophageal cancer among Singapore Chinese (DeJong, Breslow, Hong, Sridharan, and Shanmugaratnam 1974). This was a typical hospital-based interview study, with two control groups, that focused on ethnicity, diet, alcohol, and tobacco as possible risk factors. Of particular interest were questions relating to the temperature at which various beverages were consumed. We were well aware that differential "recall bias" in the interview responses of cases and controls was a strong possibility for this item. D. R. Cox's (1970) text covering logistic regression had recently appeared. Having had some previous experience with this methodology in a clinical setting (Breslow and McCann 1971), I jumped at the chance to apply the technique to the case-control study. In retrospect, the enthusiasm seems rather naive, because we simply ignored the apparent problems posed by the outcome-dependent sampling.

One aspect of our analysis that did bother me was its failure to account for the pair matching of controls to cases on age, gender, hospital ward (for one of the control groups), and time of diagnosis. Such matching was widely used to select "comparable" controls, but there was little appreci-

Table 3. Properties of Weighted (W) and Unweighted (U) Estimators

	$n_0 = n_1 = 100$		$n_0 = n_1 = 200$	
	W	U	W	U
Bias	.14	−.20	.01	−.26
Mean squared error	.35	.21	.12	.14
Standard deviation	.45	.21	.33	.24
Coverage of 95% intervals	.89	.85	.92	.75
Coverage of 99% intervals	.97	.94	.98	.87

Source: Scott and Wild (1986). Reprinted with permission.

ation among epidemiologists for the complexities that it introduced for rigorous statistical analysis. Special procedures for matched case-control designs with binary exposures were available (Miettinen 1970; Pike and Morrow 1970), but a general treatment was lacking. The problem occupied my attention on my return to Seattle in 1974 and, with the help of colleagues and students, we developed a solution based on stratified logistic regression (Breslow, Day, Halvorsen, Prentice, and Sabai 1978).

Suppose that the population at risk is so finely stratified that each case occupies a single stratum, and that the matched controls are drawn from the same stratum as the case. With S denoting the stratum, the population model is

$$\Pr(D = 1|S = j, \mathbf{X} = \mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}\beta)}{1 + \exp(\alpha_j + \mathbf{x}\beta)}.$$

This involves a separate parameter for each matched set and allows inclusion of possible interactions between exposures and matching variables among the explanatory variables $\mathbf{x}$. Following Fisherian principles, the stratum parameters α_j are eliminated by conditioning on an appropriate ancillary statistic, in this case the unordered set of exposures for the case and controls in each stratum. Thus the conditional likelihood that the exposures $\mathbf{x}_{j0}$ are those of the case and $(\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jM})$ are those of the M controls in stratum j, as observed, given the set of $M + 1$ exposures, is proportional to

$$\frac{\Pr(\mathbf{X} = \mathbf{x}_{j0}|S = j, D = 1)}{\sum_{m=0}^{M} \Pr(\mathbf{X} = \mathbf{x}_{jm}|S = j, D = 1)} \times \prod_{m=1}^{M} \Pr(\mathbf{X} = \mathbf{x}_{jm}|S = j, D = 0)}{\times \prod_{h \neq m} \Pr(\mathbf{X} = \mathbf{x}_{jh}|S = j, D = 0)}.$$

Writing $\Pr(\mathbf{X}|D) = \Pr(D|\mathbf{X})\Pr(\mathbf{X})/\Pr(D)$ in the usual fashion, the marginal probabilities drop out, and we are left with

$$\prod_{j=1}^{J} \frac{\exp(\mathbf{x}_{j0}\beta)}{\exp(\mathbf{x}_{j0}\beta) + \sum_{m=1}^{M} \exp(\mathbf{x}_{jm}\beta)} \tag{10}$$

for inference about β. Note that the terms $\exp(\mathbf{x}_{jm}\beta)$ are the relative risks for each subject relative to someone with a standard $(\mathbf{X} = \mathbf{0})$ set of exposures. These arguments are easily generalized to situations with a variable number of controls per case, and even to matched sets with an arbitrary number of cases and controls (Breslow et al. 1978). The conditional likelihood (10) also arises from the stratified logistic regression model for a cohort study, by conditioning on the number of cases that occur in each stratum. This further strengthens the notion that one is estimating the same parameters in cohort studies and case-control studies. With pair matching, (10) is formally identical to a likelihood for simple logistic regression on the *difference* in exposures between case and control (Holford, White, and Kelsey 1978).

With this new tool in hand, I was greatly relieved to discover that the results of the Singapore study were little affected by the choice of conditional versus unconditional likelihood analysis (Table 4, columns 2 and 3). The inflation of the regression coefficients (Table 4, column 4) when

Table 4.   Regression Coefficients ± Standard Errors

Variable	Conditional: likelihood	Unconditional: single α	Unconditional: multiple α_j
Ethnic group[a]	1.27 ± .33	1.42 ± .36	1.62 ± .25
Chinese wine[b]	.51 ± .29	.54 ± .28	.68 ± .27
Cigarettes[c]	.11 ± .10	.11 ± .09	.16 ± .09
Temperature[d]	.79 ± .16	.76 ± .15	1.12 ± .15

[a] 1 = Teochew and Hokkien, 0 = Cantonese and other.
[b] 1 = Consumer, 0 = Nonconsumer.
[c] Per pack of 10.
[d] Number of beverages (0–3) drunk "burning hot."
Source: Breslow 1982.

one tries to explicitly estimate the stratum parameters α_j is somewhat greater than the factor $M/(M + 1) = 1.25$ predicted by results for a single binary exposure (Breslow 1981). It well illustrates the problems of likelihood inference with large numbers of parameters. The fact that the original analysis that ignored the matching agreed with the new, correct analysis was, of course, fortuitous and suggested that the matching variables were not strongly associated with the exposures. Unmatched analyses of matched data generally yield conservative estimates of relative risk (Armitage 1975; Breslow and Day 1980, table 7.12).

Prentice and Breslow (1978), in a paper that further clarified the conceptual foundations of the case-control study, derived the conditional likelihood (10) from failure time considerations. One starts with a large (voire infinite) population that is followed forward in time. For an individual with exposures x, the disease incidence rate at time t is specified as $\lambda(t|x) = \lambda_0(t) \exp(x\beta)$ (Cox 1972). At the time t_j of occurrence of the jth disease case, M controls are sampled at random from the population. Conditioning on the unordered set of exposures for the case and controls then leads to (10). This derivation helps to explain why, with "incidence density sampling" (Miettinen 1976) where controls are sampled at the times of occurrence of the cases, the exposure odds ratio approximates the ratio of instantaneous disease rates and thus why the odds ratio is useful even for the study of common diseases (Greenland and Thomas 1982).

Although these conditional likelihood arguments were developed in the context of sampling from an infinite population, there is no reason why they cannot be applied also to sampling from an actual finite cohort. As noted earlier, this idea was already implicit in the 1959 Mantel–Haenszel paper. Mantel (1973) explicitly proposed sampling from a defined cohort, using an independent toss of a biased coin to decide whether or not each control would be included in the final sample. Motivated by a desire to reduce the computational burden, he termed the result a "synthetic" case-control study. Thomas was the first to propose sampling from the risk sets formed during a Cox regression analysis (Liddell, McDonald, and Thomas 1977). Figure 2 is a schematic of the risk sets in a cohort study. The basic idea is to replace each of them by a reduced risk set consisting of the case and a random sample (without replacement) of the remaining risk set members. Thomas proposed using the conditional likelihood (10) for inference, which of course has exactly the same form as Cox's (1975) partial likelihood

for the original risk set. Here too the initial motivation was primarily computational. But it quickly became clear that the real value of such nested case-control sampling, as it came to be called, was for selection of individuals on whom additional data could be collected. The technique is particularly valuable when stored sera or other biological materials are available for a large cohort, but expensive laboratory assays are needed for quantitative exposure assessment.

Although the intuition underlying the nested case-control study is strong, and the use of the likelihood (10) is rendered plausible by the results for matched studies, more formal justification has taken time to develop. Oakes (1981) led the way with his derivation of (10) as a partial likelihood, but these arguments were still regarded as incomplete. Only recently have rigorous proofs appeared of the asymptotic consistency and normality of relative risks estimated by partial likelihood under nested case-control sampling (Goldstein and Langholz 1992). The most interesting of these proofs develop the theory in terms of marked point processes (Borgan, Goldstein, and Langholz in press). Besides confirming the asymptotic properties of the relative risk estimates, this approach also neatly solves the problem of how to use the nested case-control sample for estimation of the baseline cumulative incidence function.

Estimation of absolute risk functions as well as relative risk functions is in principle possible from a nested case-control sample, because one knows the sampling probabilities. If data from the full cohort are available, then the standard estimator of $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ in the Cox model is

$$\hat{\Lambda}(t; \hat{\beta}) = \sum_{t_j \leq t} \frac{1}{\sum_{l \in \mathcal{R}_j} \exp(x_l \hat{\beta})} , \qquad (11)$$

where $\mathcal{R}_j$ denotes the full risk set at the time t_j of occurrence of the jth case and $\hat{\beta}$ is the partial likelihood estimate. Suppose that $\mathcal{R}_j$ contains N_j subjects including the case and that M controls are sampled for the reduced risk set $\tilde{\mathcal{R}}_j$. Borgan and Langholz (1993) and Borgan et al. (in
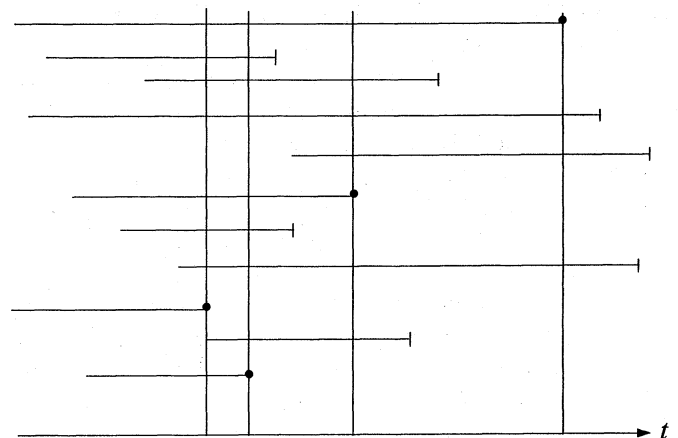

Figure 2.   Definition of Risk Sets. Each horizontal line (———) denotes the observation period for a single subject as a function of time or age. Lines that terminate in a bullet (•) correspond to cases diagnosed at that time, whereas those that terminate with a bar (|) are noncases. The risk sets defined at each time of diagnosis contain those subjects whose observation period intersects the corresponding vertical line. (Adapted from Langholz and Clayton 1994, Fig. 1).
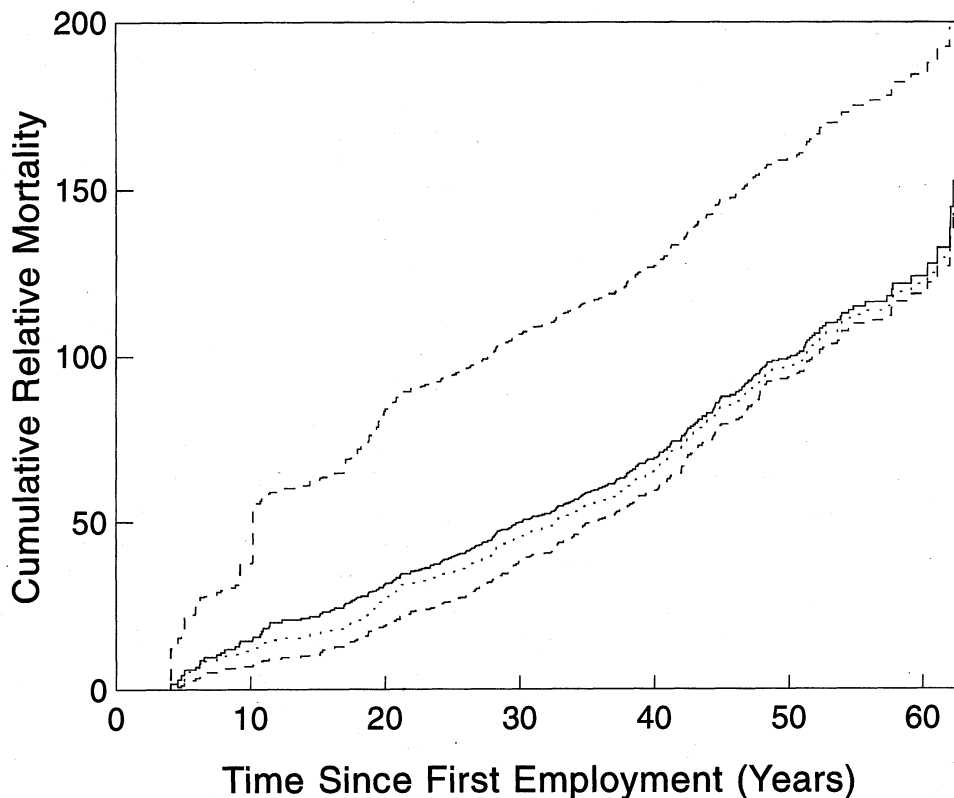
Figure 3. Estimates of the Cumulative Relative Mortality Function: M = 1 (– · –); M = 5 (· ·); M = 100 (——); Breslow and Langholz crude estimator, M = 5 (– – –). (Source: Borgan and Langholz 1993; reprinted with permission of the International Biometric Society.)

press) proposed the estimator

$$\hat{\Lambda}(t; \hat{\beta}) = \sum_{t_j \leq t} \frac{1}{\frac{N_j}{M+1} \sum_{l \in \tilde{\mathcal{R}}_j} \exp(\mathbf{x}_l \hat{\beta})} , \qquad (12)$$

which weights the contribution from $\tilde{\mathcal{R}}_j$ by the inverse ratio of its size to that of the full risk set. They showed using martingale theory that (12) is essentially unbiased, just as (11) is. Figure 3 shows the cumulative (relative) lung cancer mortality function for the Montana smelter workers considered by Breslow and Langholz (1987), who had earlier suggested a rather badly biased analog of (12) in which the cases were removed from the reduced risk sets. The new estimator with matching ratio $M = 100$ is essentially equivalent to the full cohort estimator (11). Comparing the results for it with those for smaller values of M confirms that a practical methodology now exists for estimating baseline cumulative hazard functions from nested case-control studies.

Prentice (1986) introduced the case-cohort design as an alternative method of sampling from a defined cohort. Here a random subcohort is sampled from the entire cohort at the outset of the follow-up period, and exposure information is processed for members of the subcohort. Cases occuring outside the subcohort are ascertained, and their exposure information is processed at the time of diagnosis. This design offers substantial advantages over nested case-control sampling in situations where multiple disease endpoints are to be evaluated, because unaffected members of the single subcohort may serve as controls for the disease cases

of each type. The choice between the two designs is less clear when only a single disease endpoint is to be studied. One drawback of the case-cohort design is the necessary assumption that exposure assessments, such as laboratory assays on stored biologicals, are not affected by the passage of time. Langholz and Thomas (1991) showed that the nested case-control design may have greater efficiency than the case-cohort design when there is moderate random censoring or staggered entry into the cohort. Methods of "refreshing" the subcohort so as to avoid such efficiency loss are available. (See Lin and Ying 1993 and Barlow 1994 for further discussion and methods of analysis of case-cohort data.)

6. MORE INFORMATIVE SAMPLING

Case-control studies are used to study rare diseases because they are very efficient as compared to cohort studies of the size needed to produce the same number of cases. The method concentrates resources where there is the greatest amount of information, namely on the cases. Further improvements in efficiency are possible by utilizing more complex sampling schemes to maximize the variation in the exposures of the cases and controls that are ultimately analyzed. White (1982) proposed a "two-stage" sampling design for studying the effects of a rare exposure on a rare disease, with cohort sampling at the initial stage. Breslow and Cain (1988) considered a two-stage case-control design. Table 5 illustrates the basic idea with hypothetical data from a two-stage study of lung disease and factory employment. At Stage I, 500 cases and 500 controls are drawn at ran-

Table 5. Fictitious Data From a Two-Stage Case-Control Study

Stage I (N_{ij})

Exposure	Cases	Controls
Factory workers	40	20
Employed elsewhere	460	480
Total	500	500

Stage II (n_{ijk})

	Cases ($i = 1$)		Controls ($i = 0$)	
Exposure	Factory ($j = 1$)	Elsewhere ($j = 2$)	Factory ($j = 1$)	Elsewhere ($j = 2$)
Smokers ($k = 1$)	30	37	10	42
Nonsmokers ($k = 2$)	10	133	10	128
Totals (n_{ij})	40	170	20	170

Source: White (1982). Reprinted with permission.

dom from the community and classified as to whether or not they are employed in the factory. Information on smoking is essential to adjust the relative risk estimates, but the investigators can afford to interview only about half of the 1,000 subjects. In view of the rarity of the exposure, the investigators elect to interview all of the exposed cases and controls at Stage II but only a subset of the unexposed cases and controls. This is more efficient than randomly selecting the cases and controls at Stage II without considering the information already available for them. It means, however, that the relationship between exposure and disease is distorted for those who have complete data. The statistical challenge is to combine the information available at both stages of sampling in the most efficient way so as to estimate an adjusted relative risk.

More generally, suppose that N_1 cases and N_0 controls are classified into J strata on the basis of an initial, possibly crude measure of exposure. Let N_{ij} denote the numbers of cases ($i = 1$) and controls ($i = 0$) in the jth stratum. At Stage II, n_{ij} subjects are sampled from the N_{ij} in each of the $2J$ disease × stratum cells. Additional measurements are taken for these subjects, yielding a p-dimensional vector $\mathbf{X}$ of explanatory variables that takes values $\mathbf{x}_{jk}, j = 1, \ldots, J, k = 1, \ldots K$, and that may include information obtained at either stage. Let n_{ijk} denote the number of cases or controls in stratum j with $\mathbf{X} = \mathbf{x}_{jk}$.

A key assumption is that the probability of disease development in the population depends on the stratum only through $\mathbf{X}$:

$$\Pr(D = 1 | S = j, \mathbf{X} = \mathbf{x}_{jk}) = \frac{\exp(\mathbf{x}_{jk}\beta)}{1 + \exp(\mathbf{x}_{jk}\beta)}, \quad (13)$$

where $\mathbf{x}_{jk}$ now includes the constant term. But the actual likelihood for the two-stage data is proportional to

$$\prod_{i=0}^{1} \prod_{j=1}^{J} [\Pr(S = j | D = i)]^{N_{ij}}$$

$$\times \prod_{k=1}^{K} [\Pr(\mathbf{X} = \mathbf{x}_{jk} | D = i, S = j)]^{n_{ijk}} \quad (14)$$

(Holubkov 1995; Scott and Wild 1991). With $\delta_j = \log[\Pr(S = j | D = 1)/\Pr(S = j | D = 0)]$, define

$$P_{1j} = 1 - P_{0j} = \frac{N_1 \exp(\delta_j)}{N_0 + N_1 \exp(\delta_j)}$$

to be the probability that a Stage I subject in stratum j is a case, and similarly define

$$p_{1jk} = 1 - p_{0jk} = \frac{n_{1j}\pi_0 \exp(-\delta_j + \mathbf{x}_{jk}\beta)}{n_{0j}\pi_1 + n_{1j}\pi_0 \exp(-\delta_j + \mathbf{x}_{jk}\beta)}$$

to be the probability that a Stage II subject in stratum j with exposures $\mathbf{X} = \mathbf{x}_{jk}$ is a case.

These definitions allow one to factor the likelihood (14) into terms involving (δ, β) and terms involving the marginal stratum and exposure distributions, just as in Section 4. But with the more complicated design, the unconstrained solution does not generally satisfy the constraints, namely that the marginal probabilities of being a case or a control in stratum j at Stage II are fixed by design (Schill, Jöckel, Drescher, and Timm 1993). Breslow and Holubkov (1995) concentrated the Lagrangian that arises from the constrained estimation problem and reduced it to the solution of the $(p + J)$ equations

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \left\{ n_{1jk} - \frac{(n_{1j} - T_j)n_{0j}n_{+jk}p_{1jk}}{n_{0j}n_{1j} - T_j(n_{0j} - n_{+j}p_{0jk})} \right\} \mathbf{x}_{jk} = 0, \quad (15)$$

and for $j = 1, \ldots, J$,

$$T_j - n_{1j} + \sum_{k=1}^{K} \frac{(n_{1j} - T_j)n_{0j}n_{+jk}p_{1jk}}{n_{0j}n_{1j} - T_j(n_{0j} - n_{+j}p_{0jk})} = 0, \quad (16)$$

where $T_j = N_{1j} - N_{+j}P_{1j}$. A linearized, asymptotically equivalent set of equations leads to a more easily computed variance than that given by Aitchison and Silvey (1958) for constrained likelihood estimation and facilitates demonstration of asymptotic normality even when the exposures are continuous. Less efficient pseudolikelihood estimates result

if one replaces (16) by $T_j = 0$, sets T_j to 0 in (15), and solves the resulting equations either separately (Breslow and Cain 1988) or jointly (Schill et al. 1993), this latter solution resulting from unconstrained maximization of (14).

Two-stage designs have many potential applications. One is to case-control studies with validation subsampling to correct for measurement error (Carroll, Gail, and Lubin 1993). In this case S denotes a surrogate available for all subjects, whereas X denotes error-free exposures measured for the validation subsample. The assumption of nondifferential measurement error needed to justify calling S a surrogate, namely $\Pr(S = j|X = x_{jk}, D = 0) = \Pr(S = j|X = x_{jk}, D = 1)$, is equivalent to the assumption (13) of conditional independence between D and S given X. Thus Equations (15) and (16) provide a fully efficient solution to the measurement error problem when the surrogate is discrete. Carroll et al. (1993) developed a pseudo-likelihood approach for the general problem of validation subsampling that allows for the possibility of differential measurement error and eliminates the requirement for discrete data at the first stage. However, this approach requires specification of a parametric model for the distribution of the error-prone measurements S given X and D. Robins, Rotnitzky, and Zhao (1994) provided the basic machinery needed to develop semiparametric efficient estimators for many of these problems.

An ingenious analog of the two-stage design that is applicable with nested case-control sampling was recently devised by Langholz and Borgan (1995). Here the object is to use the Stage I (stratum) information available for all members of the risk set to select those members for the reduced risk set. Because estimation of relative risk depends on the contrast in exposures between cases and controls, efficiency is enhanced by selecting them to be as different as possible vis-a-vis the exposures. Changing notation, suppose now that N_{ij} members of the risk set $\mathcal{R}(t_i)$ are in stratum j. Then one deliberately chooses the controls so that exactly M_{ij} members of the reduced risk set $\tilde{\mathcal{R}}(t_i)$ are in stratum j. When there are only two strata and a single control per case, so that $M_{i1} = M_{i2} = 1$, this means sampling the control from the *opposite* stratum as the case. To correct for the biased sampling, one weights the partial likelihood contributions of case and control by the ratio N_{ij}/M_{ij}, depending on their stratum. In practice this is easily accomplished by including $\log(N_{ij}/M_{ij})$ as an *offset* in the regression equation. The procedure is called counter-matching.

7. LIMITATIONS AND CHALLENGES

Statistics has gained a place of modest usefulness in medical research. It can deserve and retain this only by complete impartiality, which is not unattainable by rational minds ...I do not relish the prospect of this science being now discredited by a catastrophic and conspicuous howler. For it will be as clear in retrospect, as it is now in logic, that the data so far do not warrant the conclusions based on them. (Fisher 1957b, p. 298, on smoking and lung cancer.)

Statisticians have contributed immensely to the development of the modern case-control study. Their conceptualization in terms of sampling from a fictitious or actual cohort study illuminates the close relationship between these two principal methodologies of analytic epidemiology. The role of matching in study design, and methods to account for it in the analysis, are now much better understood. Logistic and other relative risk regression procedures based on likelihood concepts provide epidemiologists with modern statistical tools for model validation and outlier detection (Hosmer and Lemeshow 1989). Semiparametric methods such as the generalized additive model (Hastie and Tibshirani 1990) allow them to visualize their data in new ways, leading to new insights and hypotheses.

But despite these technical advances, the fundamental problems of drawing causal inferences from observational data persist. Fisher remained skeptical of the claim that cigarette smoking caused lung cancer even in the face of what the medical community regarded as overwhelming evidence. His extreme viewpoint is best understood by recalling that he was both the geneticist, well aware of the influence of heredity on disease, and the statistician, who had perfected randomization as *the* method of drawing causal conclusions in experimental settings. He also took sharp exception to what he regarded as the hysterical reaction of the public media to an unproven hypothesis.

I hope that most of us, as statisticians with rational minds, agree today that Fisher was seriously mistaken about the hazards of cigarette smoking. Regrettably, his prominent position on this issue may have helped delay much-needed educational programs and regulatory action. In this country, lung cancer surpassed breast cancer as a cause of mortality in women some 7–8 years ago. Fisher's observation of low lung cancer rates among smoking women in the mid-1950s simply reflected a delay in the epidemic due the fact that women started smoking later than men and that decades were required for the multistep carcinogenic process to start affecting women in large numbers. His emphasis on the genotype was relevant, not as a cause of lung cancer per se, but rather through *interactions* with cigarette smoking. For example, evidence is accumulating that individuals with high oxidative or low detoxification capacity vis-a-vis known carcinogens in cigarette smoke, due to genetic polymorphisms in well-characterized enzyme systems, may be at particularly high risk (Kawajiri, Nakachi, Imai, Watanabe, and Hayashi 1993; Nazar-Stewart et al. 1993).

Although Fisher's position on this particular association was wrong, his concern about negative public reaction to dubious scientific claims was well founded. The public is increasingly weary and skeptical of the multitude of contradictory reports of health hazards emanating from epidemiology. A recent *Science* news article noted that, although epidemiologists themselves generally discount isolated reports of weak associations, with relative risks under two or three, such studies are published by the "journal load" and many receive prominent media coverage (Taubes 1995). Given the pressure on young faculty at universities and research institutes to publish, and given a funding mechanism that discourages large, long-term team projects, the plethora of "false alarms" will undoubtedly continue.

Disease Outcome   Latent Exposure (s)   Measurements
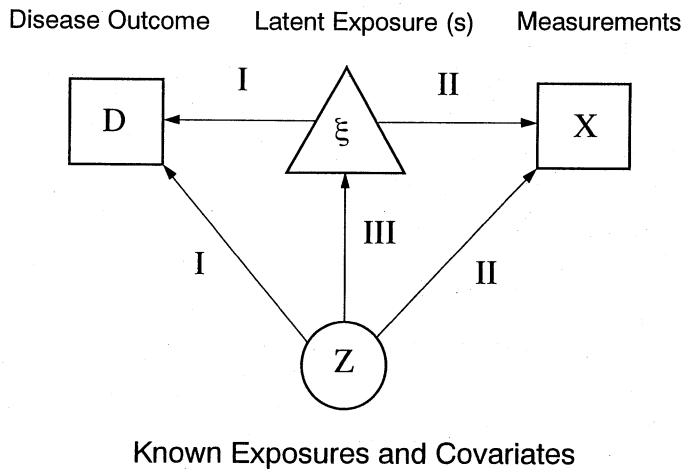

Known Exposures and Covariates

Figure 4. A General Model for Measurement Error. The labels I, II, and III denote submodels (see text). (©1992 Oxford University Press. Reprinted with permission.)

Figure 5. A Graphical Model of a Randomized Intervention (Instrumental Variable) Study. U represents both observed and unobserved factors that may confound the association of actual treatment and response. (©1995 Elsevier Science BV. Reprinted with permission.)

The limitations of case-control methodology are well known (Austin, Hill, Flanders, and Greenberg 1994). They include selection bias, most often caused by high rates of nonparticipation that render the controls nonrepresentative of the population at risk; measurement error, particularly differential "recall bias" among cases and controls; and confounding, the ever-present possibility that the observed association is the result of "hidden variables" that embody the true causal relationships. The random-digit dialing procedure widely used for control selection may often yield a control sample in which the lower socioeconomic strata are underrepresented. Despite Dorn's (1959) recommendation for objective exposure assessment, the overwhelming majority of case-control studies continue to rely on a questionnaire as the primary data collection instrument. For example, of the 223 studies published in 1992 that were mentioned in Section 1, 150 (67.4%) used a questionnaire as the only source of exposure data and another 35 (15.7%) combined questionnaire responses with more objective measurements (Correa et al. 1994). Although the bias toward the null that typically affects relative risk estimates in the presence of nondifferential error in the measurement of exposures is well known, the fact that measurement error in potential confounders can render them useless for adjustment purposes is not so well known (Greenland and Robins 1985).

What can and should we as statisticians do to overcome these difficulties? I was quoted in the previously mentioned Science article (Taubes 1995) as saying that multiple regression analyses, of the type to which much of this article has been devoted, are of little help in solving these problems and may even give some investigators a false sense of security that they have done so. Regression analysis works well for adjusting relative risk estimates for the effects of known confounders that are measured without error, but reality often lies elsewhere. More complex models and analyses that account for the measurement error are required. The most hopeful situation is where error-free measurements are available for some subjects, as in a validation substudy. The efforts of Carroll and Robins and their colleagues have
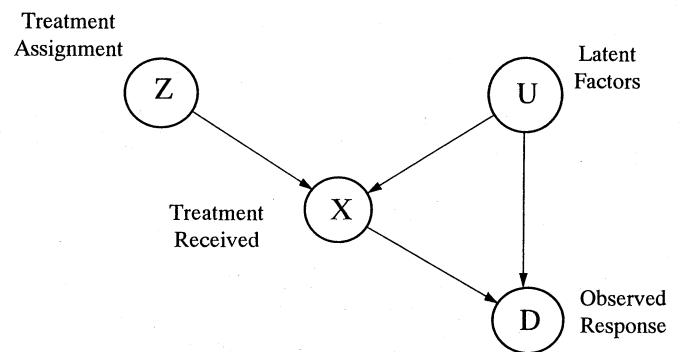
been mentioned already as examples of how to proceed in this case. Many others could be cited. When there is no "gold standard," but replicate measurements with independent errors are available for some subjects, structural equation approaches are available (Clayton 1992; Plummer and Clayton 1993). Figure 4 shows one scenario requiring specification of three different models: I, a model for disease D as a function of the true exposures, whether (Z) or not (ξ), directly measurable; II, a model for error-prone measurements X as a function of true exposures ξ and covariables Z; and III, an exposure model that specifies the distribution of ξ as a function of Z. Of course, such models typically involve strong parametric assumptions that need to be evaluated carefully.

Finally, what can we do in the face of the association versus causation dilemma? My own position on this issue is close to what has been termed the "classical epidemiology school." One considers the available epidemiologic data in conjunction with those from whole animal experiments and the laboratory in an attempt to arrive informally at a judgment as to whether the association is causal. Bradford Hill's (1971) criteria regarding the strength of the association, consistency in different settings, presence of a dose–response trend, timing, biological plausibility, and so on are relevant to this approach. One must admit, however, that these criteria have not been as successful in sorting out the signal from the noise as one might have hoped some 30 years ago, in part because they have not been applied uniformly. If we are to make further progress, other approaches may be needed.

Today, the randomized intervention or prevention trial is very much in vogue as a method of demonstrating risk factor effects. Due to the limited variation in many life-style factors within individual cultures, there is little opportunity with case-control or cohort studies to observe relative risks of sufficient magnitude to overcome the measurement error and confounding biases. The proposed solution is randomized encouragement to change unhealthy behaviors, or randomization of preventive medications. Such trials can be very effective when they work, in the sense that there are clear differences in outcome between the randomized treatment groups. Interpretation of a negative result is much

more difficult. Furthermore, because the costs of such trials as the Multiple Risk Factor Intervention Trial (MRFIT) and the Women's Health Initiative (WHI) run into the hundreds of millions of dollars, society can afford only a limited number of them.

Another school of thought, coming from philosophy, computer science, social science, and even statistics, argues that randomization has been oversold as a means of evaluating social intervention programs. Protagonists hold that we can get clearer answers from the available observational data by thinking hard about causal relationships among variables and by integrating our knowledge of causal structures into the data analysis. Examples include the econometrician's instrumental variable analysis, Rubin's (1974) causal model, and Robins' (1986) G-computation algorithm for longitudinal data. Pearl (1995a) has developed procedures for causal reasoning based on directed, acyclic graphs in a paper that synthesizes and explicates much earlier work. His graphical approach to an instrumental variable analysis is shown in Figure 5 (Pearl 1995b). Here Z, the instrumental variable, is marginally independent of the unobserved confounding factors U and influences the disease outcome D only through its effect on X, the exposure or treatment variable under study.

I find it intriguing that the paradigm for an instrumental variable, as in Figure 5, is in fact a randomized intervention and that X represents the treatment actually applied to or accepted by the subject. This suggests that, at present, the most important use of causal analysis may lie in the interpretation of results from randomized intervention trials that have substantial noncompliance. Whether or not such methods will prove useful with purely observational studies depends on the identification of credible instruments and is still an open question.

Interest in graphical models that accommodate unobserved or latent variables as a means of approaching such difficult issues as measurement error and confounding bias is currently very high. Resolution of these issues, whether by graphical modeling, randomized intervention trials, or more classical analyses of observational data, will challenge statistical research workers for many years to come.

REFERENCES

Aitchison, J., and Silvey, S. D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Restraints," *Annals of Mathematical Statistics*, 29, 813–828.

Anderson, J. A. (1972), "Separate Sample Logistic Discrimination," *Biometrika*, 59, 19–35.

Armenian, H. K., and Lilienfeld, D. E. (1994), "Overview and Historical Perspective," *Epidemiologic Reviews*, 16, 1–5.

Armitage, P. (1975), "The Use of the Cross-Ratio in Aetiological Surveys," in *Perspectives in Probability and Statistics*, ed. J. Gani, London, Academic Press, pp. 349–355.

Austin, H., Hill, H. A., Flanders, W. D., and Greenberg, R. S. (1994), "Limitations in the Application of Case-Control Methodology," *Epidemiologic Reviews*, 16, 65–76.

Barlow, W. E. (1994), "Robust Variance Estimation for the Case-Cohort Design," *Biometrics*, 50, 1064–1072.

Borgan, O., Goldstein, L., and Langholz, B. (in press), "Methods for the Analysis of Sampled Cohort Data in the Cox Proportional Hazards Model," *The Annals of Statistics*.

Borgan, O., and Langholz, B. (1993), "Nonparametric Estimation of Relative Mortality From Nested Case-Control Studies," *Biometrics*, 49, 593–602.

Breslow, N. (1976), "Regression Analysis of the Log Odds Ratio: A Method for Retrospective Studies," *Biometrics*, 32, 409–416.

——— (1981), "Odds Ratio Estimators When the Data are Sparse," *Biometrika*, 68, 73–84.

——— (1982), "Covariance Adjustment of Relative-Risk Estimates in Matched Studies," *Biometrics*, 38, 661–672.

Breslow, N. E., and Cain, K. C. (1988), "Logistic Regression for Two-Stage Case-Control Data," *Biometrika*, 75, 11–20.

Breslow, N. E., and Clayton, D. C. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E., and Day, N. E. (1980), *Statistical Methods In Cancer Research I. The Analysis of Case-Control Studies*, Lyon, France: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978), "Estimation of Multiple Relative Risk Functions in Matched Case-Control Studies," *American Journal of Epidemiology*, 108, 299–307.

Breslow, N. E., and Holubkov, R. (1995), "Fitting Logistic Regression Models to Data From Two-Stage Case-Control Studies," Technical Report 136, University of Washington, Dept. of Biostatistics.

Breslow, N., and Langholz, B. (1987), "Nonparametric Estimation of Relative Mortality Functions," *Journal of Chronic Diseases*, 40 (Suppl. 2), 89S–99S.

Breslow, N., and McCann, B. (1971), "Statistical Estimation of Prognosis for Children With Neuroblastoma," *Cancer Research*, 31, 2098–2103.

Broders, A. C. (1920), "Squamous-Cell Epithelioma of the Lip. A Study of Five Hundred and Thirty-Seven Cases," *Journal of the American Medical Association*, 74, 656–664.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993), "Case-Control Studies With Errors in Covariates," *Journal of the American Statistical Association*, 88, 185–199.

Carroll, R. J., Wang, S., and Wang, C. Y. (1995), "Prospective Analysis of Logistic Case-Control Studies," *Journal of the American Statistical Association*, 90, 157–169.

Clayton, D. (1992), "Models for the Longitudinal Analysis of Cohort and Case-Control Studies with Inaccurately Measured Exposures. Application: Dietary Ratio of Polyunsaturated to Saturated Fats and Subsequent Risk of Coronary Heart Disease," in *Statistical Models for Longitudinal Studies of Health*, eds. J. H. Dwyer, F. Manning, P. Lippert, and H. Hoffmeister, New York: Oxford University Press, pp. 301–331.

Clayton, D., and Hills, M. (1993), *Statistical Models in Epidemiology*, New York: Oxford University Press.

Cochran, W. G. (1954), "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451.

Cole, P. (1979), "The Evolving Case-Control Study," *Journal of Chronic Diseases*, 32, 15–27.

Cologne, J. B., and Breslow, N. E. (1990), "A Small Sample Simulation Study of Methods for Log Odds Ratio Regression Analysis," *Computational Statistics & Data Analysis*, 9, 217–235.

Cornfield, J. (1951), "A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix," *Journal of the National Cancer Institute*, 11, 1269–1275.

——— (1956), "A Statistical Problem Arising from Retrospective Studies," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley, CA: University of California Press, pp. 135–148.

Cornfield, J., Gordon, T., and Smith, W. W. (1961), "Quantal Response Curves for Experimentally Uncontrolled Variables," *Bulletin of the International Statistical Institute*, 38, 97–115.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.

Correa, A., Stewart, W. F., Yeh, H. C., and Santos-Burgoa, C. (1994), "Exposure Measurement in Case-Control Studies: Reported Methods

and Recommendations," *Epidemiologic Reviews*, 16, 18–32.

Cosslett, S. R. (1981), "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49, 1289–1316.

Cox, D. R. (1966), "Some Procedures Connected With the Logistic Qualitative Response Curve," in *Research Papers in Statistics: Festschrift for J. Neyman*, ed. F. N. David, New York: John Wiley, pp. 55–71.

——— (1970), *Analysis of Binary Data*, London: Chapman and Hall.

——— (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Cytel Software Corporation, (1995), *StatXact-3 for Windows User Manual*, Cambridge, MA: Author.

Davis, L. J. (1985), "Generalization of the Mantel–Haenszel Estimator to Non-Constant Odds Ratio," *Biometrics*, 51, 437–495.

Day, N. E., and Kerridge, D. F. (1967), "A General Maximum Likelihood Discriminant," *Biometrics*, 23, 313–323.

DeJong, U. W., Breslow, N., Hong, G. E., Sridharan, M., and Shanmugaratnam, K. (1974), "Etiological Factors in Oesophageal Cancer in Singapore Chinese," *International Journal of Cancer*, 13, 291–303.

Donner, A., and Hauck, W. (1988), "Estimation of a Common Odds Ratio in Case-Control Studies of Familial Aggregation," *Biometrics*, 44, 369–378.

Dorn, H. F. (1959), "Some Problems Arising in Prospective and Retrospective Studies of the Etiology of Disease," *The New England Journal of Medicine*, 261, 571–579.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, Ser. A, 222, 309–368.

——— (1935), "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, 98, 39–54.

——— (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

——— (1957a), "Dangers of Cigarette Smoking," *British Medical Journal*, July 6, 43.

——— (1957b), "Dangers of Cigarette Smoking," *British Medical Journal*, August 3, 297–298.

——— (1958a), "Lung Cancer and Cigarettes," *Nature*, 182, 108.

——— (1958b), "Cancer and Smoking," *Nature*, 182, 596.

Gart, J. M. (1970), "Point and Interval Estimation of the Common Odds Ratio in the Combination of 2 × 2 Tables With Fixed Marginals," *Biometrika*, 57, 471–475.

Goldstein, L., and Langholz, B. (1992), "Asymptotic Theory for Nested Case-Control Sampling in the Cox Regression Model," *The Annals of Statistics*, 20, 1903–1928.

Greenland, S., and Robins, J. M. (1985), "Confounding and Misclassification," *American Journal of Epidemiology*, 122, 495–506.

Greenland, S., and Thomas, D. C. (1982), "On the Need for the Rare Disease Assumption in Case-Control Studies," *American Journal of Epidemiology*, 116, 547–553.

Harkness, W. L. (1965), "Properties of the Extended Hypergeometric Distribution," *Annals of Mathematical Statistics*, 36, 938–945.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Hauck, W. (1979), "The Large Sample Variance of the Mantel–Haenszel Estimator of a Common Odds Ratio," *Biometrics*, 35, 817–819.

Hill, A. B. (1971), *Principles of Medical Statistics* (9th ed.), New York: Oxford University Press.

Holford, T. R., White, C., and Kelsey, J. L. (1978), "Multivariate Analysis for Matched Case-Control Studies," *American Journal of Epidemiology*, 107, 245–256.

Holubkov, R. (1995), *Maximum Likelihood Estimation in Two-Stage Case-Control Studies*, Ph.D. dissertation, University of Washington, Dept. of Biostatistics.

Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley, CA: University of California Press, pp. 221–233.

Jones, M. P., O'Gorman, T. W., Lemke, J. H., and Woolson, R. F. (1989),

"A Monte Carlo Investigation of Homogeneity Tests of the Odds Ratio Under Various Sample Size Configurations," *Biometrics*, 45, 171–181.

Kahn, H. A., and Sempos, C. T. (1989), *Statistical Methods in Epidemiology*, New York: Oxford University Press.

Kawajiri, K., Nakachi, K., Imai, K., Watanabe, J., and Hayashi, S. (1993), "The CYP1A1 Gene and Cancer Susceptibility," *Critical Reviews In Oncology/Hematology*, 14, 77–87.

Kneale, G. (1971), "Problems Arising in Estimating From Retrospective Survey Data the Latent Periods of Juvenile Cancers Initiated by Obstetric Radiography," *Biometrics*, 27, 563–590.

Lane-Claypon, J. E. (1926), "A Further Report on Cancer of the Breast," *Reports on Public Health and Medical Subjects*, 32, Ministry of Health, London: Her Majesty's Stationery Office.

Langholz, B., and Borgan, O. (1995), "Counter-Matching: A Stratified Nested Case-Control Sampling Method," *Biometrika*, 82, 1–11.

Langholz, B., and Clayton, D. (1994), "Sampling Strategies in Nested Case-Control Studies," *Environmental Health Perspectives*, 102, 47–51.

Langholz, B., and Thomas, D. C. (1991), "Efficiency of Cohort Sampling Designs: Some Surprising Results," *Biometrics*, 47, 1563–1571.

Liang, K. Y. (1985), "Odds Ratio Inference With Dependent Data," *Biometrika*, 72, 678–682.

——— (1987), "Extended Mantel–Haenszel Estimating Procedure for Multivariate Logistic Regression Models," *Biometrics*, 43, 289–299.

Liang, K. Y., Beaty, T. H., and Cohen, B. H. (1986), "Application of Odds Ratio Regression Models for Assessing Familial Aggregation From Case-Control Studies," *American Journal of Epidemiology*, 124, 678–683.

Liang, K. Y., and Self, S. G. (1985), "Tests for Homogeneity of Odds Ratio When the Data are Sparse," *Biometrika*, 72, 353–358.

Liddell, J. R., McDonald, J. C., and Thomas, D. C. (1977), "Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining," *Journal of the Royal Statistical Society*, Ser. A, 140, 469–491.

Lilienfeld, A. M., and Lilienfeld, D. E. (1979), "A Century of Case-Control Studies: Progress?," *Journal of Chronic Diseases*, 32, 5–13.

Lin, D. Y., and Ying, Z. (1993), "Cox Regression With Incomplete Covariate Measurements," *Journal of the American Statistical Association*, 88, 1341–1349.

Lombard, H. L., and Doering, C. R. (1928), "Cancer Studies in Massachusetts. 2. Habits, Characteristics and Environment of Individuals With and Without Cancer," *New England Journal of Medicine*, 198, 481–487.

Manski, C. F., and Lerman, S. R. (1977), "The Estimation of Choice Probabilities From Choice Based Samples," *Econometrica*, 45, 1977–1988.

Manski, C. F., and McFadden, D. (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. F. Manski and D. McFadden, Cambridge, MA: MIT Press, pp. 2–50.

Mantel, N. (1973), "Synthetic Retrospective Studies and Related Topics," *Biometrics*, 29, 479–486.

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719–748.

Mehta, C. R., Patel, N. R., and Gray, R. (1985), "Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 × 2 Contingency Tables," *Journal of the American Statistical Association*, 80, 969–973.

Miettinen, O. S. (1970), "Estimation of Relative Risk From Individually Matched Series," *Biometrics*, 26, 75–86.

——— (1972), "Components of the Crude Risk Ratio," *American Journal of Epidemiology*, 96, 168–172.

——— (1976), "Estimability and Estimation in Case-Referent Studies," *American Journal of Epidemiology*, 103, 226–235.

Nazar-Stewart, V., Motulsky, A. G., Eaton, D. L., White, E., Hornung, S. K., Leng, Z. T., Stapleton, P., and Weiss, N. S. (1993), "The Glutathione S-Transferase Polymorphism as a Marker for Susceptibility to Lung Carcinoma," *Cancer Research*, 53, 2313–2318.

Neyman, J. (1955), "Statistics—Servant Of All Sciences," *Science*, 122, 401–406.

——— (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," *Probability and Statistics*, ed. U. Grenander, New York: John Wiley, pp. 213–234.

Oakes, D. (1981), "Survival Times: Aspects of Partial Likelihood," *International Statistical Review*, 49, 235–264.

Pearl, J. (1995a), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–709.

——— (1995b), "Causal Inference From Indirect Experiments," *Artificial Intelligence in Medicine*, 7, 561–582.

Phillips, A., and Holland, P. W. (1987), "Estimators of the Variance of the Mantel–Haenszel Log-Odds-Ratio Estimate," *Biometrics*, 43, 425–431.

Pike, M. C., and Morrow, R. H. (1970), "Statistical Analysis of Patient-Control Studies in Epidemiology: Factor Under Investigation an All-Or-None Variable," *British Journal of Preventive & Social Medicine*, 24, 42–44.

Plummer, M., and Clayton, D. (1993), "Measurement Error in Dietary Assessment: An Investigation Using Covariance Structure Models. Part I," *Statistics in Medicine*, 12, 925–935.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

Prentice, R. L., and Breslow, N. E. (1978), "Retrospective Studies and Failure Time Models," *Biometrika*, 65, 153–158.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

Robins, J. M. (1986), "A New Approach to Causal Inference in Mortality Studies With a Sustained Exposure Period—Applications to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J., Breslow, N., and Greenland, S. (1986), "Estimators of the Mantel–Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, 42, 311–323.

Robins, J., Greenland, S., and Breslow, N. (1986), "A General Estimator for the Variance of the Mantel–Haenszel Odds Ratio," *American Journal of Epidemiology*, 124, 719–723.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed,"

Journal of the American Statistical Association, 89, 846–866.

Rothman, K. J. (1986), *Modern Epidemiology*, Boston: Little, Brown.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Schlesselman, J. J. (1978), "Assessing Effects of Confounding Variables," *American Journal of Epidemiology*, 108, 3–8.

Schill, W., Jöckel, K. H., Drescher, K., and Timm, J. (1993), "Logistic Analysis in Case-Control Studies Under Validation Sampling," *Biometrika*, 80, 339–352.

Scott, A. J., and Wild, C. J. (1986), "Fitting Logistic Models Under Case-Control or Choice Based Sampling," *Journal of the Royal Statistical Society*, Ser. B, 48, 170–182.

——— (1991), "Fitting Logistic Regression Models in Stratified Case-Control Studies," *Biometrics*, 47, 497–510.

Seigel, D. G., and Greenhouse, S. W. (1973), "Multiple Relative Risk Functions in Case-Control Studies," *American Journal of Epidemiology*, 97, 324–331.

Stewart, A., and Kneale, G. W. (1970), "Age-Distribution of Cancers Caused by Obstetric X-Rays and Their Relevance to Cancer Latent Periods," *Lancet*, ii, 4–8.

Surgeon General (1964), *Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service*, Washington, DC: U.S. Department of Health, Education, and Welfare.

Taubes, G. (1995), "Epidemiology Faces Its Limits," *Science*, 269, 164–169.

Tarone, R. E. (1985), "On Heterogeneity Tests Based on Efficient Scores," *Biometrika*, 72, 91–95.

White, J. E. (1982), "A Two Stage Design for the Study of the Relationship Between a Rare Exposure and a Rare Disease," *American Journal of Epidemiology*, 115, 119–128.

Xie, Y., and Manski, C. F. (1989), "The Logit Model and Response-Based Samples," *Sociological Methods & Research*, 17, 283–302.