

ANALYSIS OF PROPORTIONATE MORTALITY DATA USING LOGISTIC REGRESSION MODELS

JAMES M. ROBINS¹ AND DON BLEVINS¹

Robins, J. M. (Harvard School of Public Health, Boston, MA 02115) and D. Blevins, Analysis of proportionate mortality data using logistic regression models. *Am J Epidemiol* 1987;125:524-35.

When only proportionate mortality data are available to an investigator studying the effect of an exposure on a particular cause of death, controls must be selected from among persons dying of other causes believed to be uninfluenced by the exposure under study. When qualitative or quantitative estimates of exposure history can be obtained for the deceased individuals, it is shown that one can use logistic regression models for the mortality odds to efficiently estimate the effect of exposure while controlling for relevant confounding factors by incorporating a priori information on baseline mortality rates available from US life tables. The proposed method is used to reanalyze data from a cohort of arsenic-exposed workers in a Montana copper smelter.

epidemiologic methods; models, theoretical; mortality; retrospective studies

In epidemiologic investigations of work-related mortality, often only deceased persons are identified. Given such proportionate mortality data, an investigator interested in the effect of an exposure of interest on a particular cause of death D_1 will choose controls from among cohort members dying of one or more causes D_2 which are believed to be uninfluenced by the exposure under study. Miettinen and Wang (1) have shown that if death from D_2 is uninfluenced by the exposure under study, then the mortality odds ratio param-

eter, but not the proportionate mortality ratio parameter, equals the standardized mortality ratio parameter. In this paper we generalize these results. Following Prentice et al. (2), we demonstrate how parameters of multivariate models describing the dependence of mortality from cause of death D_1 on an exposure of interest and on various confounding factors can be efficiently estimated from proportionate mortality data using (unconditional) logistic regression.

Suppose by means of interview and company records, work histories and qualitative or quantitative estimates of exposure history can be obtained for the deceased individuals. An internal (dose response) comparison is then feasible. An internal comparison will not be subject to the healthy worker bias associated with a standard mortality odds ratio analysis in which the age-calendar year-specific mortality odds of the cohort is compared with that of the general US population. Prentice and Breslow (3) and Breslow et al. (4) have shown that when exposure is not a cause of death from D_2 , a case-control analysis matching

Received for publication August 26, 1985, and in final form July 25, 1986.

¹ Occupational Health Program, Harvard School of Public Health, 665 Huntington Ave., Boston, MA 02115. (Reprint requests to Dr. James M. Robins.)

Dr. Robins' research was funded in part by Public Health Service grants no. 5 R23 ES03045, no. 5 P30 ES00002, and grants from the American Lung Association and the American Heart Association.

The authors acknowledge the independent work by Dr. Norman Breslow of the University of Washington School of Public Health; and the work jointly by Dr. William J. Butler, University of Michigan School of Public Health, and Robert M. Park, Health and Safety Department, UAW International Union, on the same topic that is presented in this paper.

on both age and calendar year of death can yield valid estimates of the relative risk (rate ratio) for the effect of exposure level on death from D_1 . Unfortunately, if the cohort under study is small in size, a matched analysis of proportionate mortality data can be highly inefficient. For example, suppose 20–40 deaths from cause D_1 and only twice that number from cause D_2 are available. If one matches rather precisely on age and calendar period, many cases and controls may remain unmatched. Therefore, matching is typically done on rather broad categories of age and calendar period. If the ratio of the age-specific incidence of death from D_1 to that of death from D_2 is changing rapidly with age and if the biologically relevant exposure index is highly associated with age, then broad category matching could lead to significant intrastratum confounding by age. When the biologically relevant exposure index is cumulative exposure (possibly lagged some number of years to allow for a biologic latent period), this exposure index would commonly, although not invariably, be strongly associated with age.

An alternative to matching on broad categories of age would be to enter age as an independent variable in a logistic regression model for the mortality odds. But if the effect of age on the mortality odds is not truly linear on a logistic scale (i.e., the model is misspecified), and age is associated with the exposure index, the estimate of the exposure effect will be biased. If the investigator decides to enter both a linear and a quadratic term for age in the logistic model in order to guard against bias due to model misspecification, efficiency may suffer because several age parameters must be estimated from a small sample, and the standard error for the exposure effect may increase. If the investigator wishes to further guard against bias by entering terms for calendar period and for age-calendar period interactions, the standard error for the exposure effect may increase even further. A standard approach to resolving this tension between bias and variance is to

adjust the exposure effect for, say, a quadratic age effect only when the coefficient for the age² term is significantly different from zero. But, due to lack of power in small samples, even if the linear logistic model for the age effect provides a good fit to the data (i.e., the coefficient for age² is not significant), the exposure effect estimate may still be biased if one does not adjust for the quadratic age effect as well (5, 6).

In order simultaneously to retain good efficiency and avoid significant bias, one would wish to use models that incorporate assumptions concerning the effect of age and calendar period on mortality that one believes are more likely to accurately reflect reality than the rather arbitrary assumption of a linear logistic age effect. As an example, suppose one assumes that for each disease type the ratio of the mortality rate in the cohort (among the unexposed) to that of the general US population is constant over age and calendar period. The particular value for the constant may vary with disease type. If less than one, the constant for a particular disease would represent a disease-specific (multiplicative) healthy worker effect constant over age and calendar period. This paper will show how one may utilize models that incorporate such a priori assumptions on background incidence rates to efficiently estimate exposure effects from small proportionate mortality studies.

Note that one may be uncertain whether the assumption of "constant disease-specific healthy worker effects" is likely to be a more accurate reflection of the true state of nature than the simple assumption of a linear logistic effect of age on the mortality odds. In such a case, we would recommend that data be analyzed under each assumption so that one may determine whether changes in (plausible) prior assumptions (i.e., model choice) would result in large changes in one's inferences about the exposure effect. That is, one should perform a sensitivity analysis. If the exposure effect estimate is insensitive to model choice,

one's confidence in the accuracy of this estimate will increase. On the other hand, if the between-model variation in the effect estimate is large and each model provides a good fit to the data, one will (appropriately) remain uncertain as to the true magnitude of the exposure effect (even when the model-specific "standard errors," routinely included in the computer output, are small).

UNCONDITIONAL LOGISTIC REGRESSION MODELS

Breslow et al. (4) have shown that it is possible to utilize a priori assumptions on background incidence rates in a full cohort analysis. Breslow et al. (4) reanalyzed data assembled by Lee and Fraumeni (7) on 8,047 arsenic-exposed white males who worked at a Montana copper smelter for at least one year between 1937 and 1956. The analysis by Breslow et al. was based on follow-up through December 31, 1963, at which point 142 deaths from lung cancer (disease D_1) and 714 deaths from cardiac disease (D_2) had occurred.

Following Breslow et al. (4), we assume that for both heart disease and lung cancer the relative risk (i.e., rate ratio) for a foreign-born person with z_1 and z_2 cumulative years of exposure to medium and heavy concentrations of arsenic, respectively, compared with an unexposed native-born cohort member is

$$\exp(\beta_{1,i}z_1 + \beta_{2,i}z_2 + \beta_{3,i}FB) \quad (1)$$

where FB is a covariate that takes the value 1 if the subject was foreign-born and takes the value 0 otherwise and $(\beta_{1,i}, \beta_{2,i}, \beta_{3,i})$ is a disease-specific column of regression coefficients (with i indexing disease type). Equation 1 contains the assumption that the relative risk is constant over age and calendar period.

Furthermore, again following Breslow et al., we assume that the ratio of the disease-age-calendar-specific mortality rate of an unexposed native-born cohort member to that of a member of the US white male population is constant over age and calendar period. To facilitate the exposition, we

shall refer to the above assumption as the assumption of disease-specific healthy worker effects constant over age and calendar period, since the disease-specific constants will often be less than 1. Nonetheless, our results remain valid even if the constant for heart disease, lung cancer or both diseases exceeds 1.

We now suppose that, in a proportionate mortality study, an investigator has available a random sample of the heart disease and lung cancer deaths in the cohort (the sampling fractions may differ for the two causes of death). (One must be sensitive to the fact that, in practice, the sample of deaths available for analysis may not be a representative sample of all the deaths occurring in the cohort.) For each lung cancer case and heart disease control, data on calendar year and age at death, z_1 and z_2 , and FB are available (where z_1 , z_2 , and FB are as defined above). Then, asymptotically unbiased and efficient estimators of the (estimable) unknown parameters are provided by fitting by the method of unconditional maximum likelihood the prospective logistic regression model

$$\ln \left[\frac{p(D_1 | s, t, z_1, z_2, FB)}{p(D_2 | s, t, z_1, z_2, FB)} \right] = \beta_0 + \ln RR(s, t) + \beta_1 z_1 + \beta_2 z_2 + \beta_3 FB \quad (2)$$

where $\beta_1 = \beta_{1,1} - \beta_{1,2}$; $\beta_2 = \beta_{2,1} - \beta_{2,2}$; and $\beta_3 = \beta_{3,1} - \beta_{3,2}$ and $\ln RR(s, t)$ is the natural logarithm of the ratio of lung cancer to cardiac mortality rates in US white males of age t in calendar year s .

From equation 2 it is clear that if cumulative exposure to arsenic was a risk factor for cardiac death (i.e., $\beta_{1,2}$ and $\beta_{2,2}$ were unknown and nonzero), then we could not estimate the parameters $\beta_{1,1}$ and $\beta_{2,1}$ which measure the effect of cumulative arsenic exposure on lung cancer mortality.

Equation 2 can be intuitively derived by considering the conditional probability that a death from lung cancer occurs at time t in calendar s with covariates z_1 , z_2 , FB , given that a death from lung cancer or heart disease has occurred with these specifications (2, 8). When the sampling fraction of

one or both of the causes of death is chosen by the investigator, fitting equation 2 by unconditional logistic regression can still be justified by the argument given in the Appendix.

To actually fit equation 2, we proceed as follows. For each subject, whether case (D_1) or control (D_2), we calculate covariates z_1 and z_2 (i.e., cumulative years of medium and high arsenic exposure up to the subject's time of death) and FB . We then compute $\ln RR(s,t)$ by looking up the age-calendar year-specific US white male death rates for lung cancer and cardiac disease appropriate for the age and calendar year of the subject's death and taking the natural log of the ratio. Thus, for each subject we have four covariates plus their case and control status. Equation 2 differs from a usual logistic regression model only in that the coefficient for the covariate $\ln RR(s,t)$ is not estimated from the data but is assumed a priori to be 1. Such constraints on parameter values can be incorporated by using the offset command in the GLIM3 system (9) or by using the SAS program NLIN (10).

A test of the a priori assumption that the disease-specific healthy worker effects are constant over age and calendar period is afforded by fitting equation 2 with a parameter, β_4 , for covariate $\ln RR(s,t)$ unconstrained. If a 95 per cent confidence interval for β_4 includes 1, the data are consistent with the prior assumption of constant healthy worker effects. An alternate test of the assumption of a constant healthy worker effect would be to add to equation 2 terms of the form $\beta_4 \text{ age} + \beta_5 \text{ calendar year}$ (11). If joint 95 per cent confidence intervals for β_4 and β_5 include 1, the data are again consistent with the assumption of constant healthy worker effects.

WORKED EXAMPLES

Example 1.

In table 1 we use a variety of different analytic approaches to estimate the parameters of equation 1 for death from lung cancer. Columns 1 and 2 of the table are

taken directly from Breslow et al. (4). Column 1 represents fitting equation 1 from the full copper smelter cohort data using the Cox proportional hazard model. The lung cancer age-calendar-specific baseline hazards (incidences) are treated as nuisance parameters in a Cox analysis. The analysis represented in column 2 also utilizes full cohort data but, in contrast to a Cox analysis, incorporates the a priori assumption that the healthy worker effect for lung cancer is constant over age and calendar period. The rows labeled "medium arsenic," "heavy arsenic," and "foreign-born" contain estimates of the coefficients $\beta_{1,1}$, $\beta_{2,1}$, and $\beta_{3,1}$. In column 3 are the estimates of β_1 , β_2 , and β_3 (defined following equation 2) obtained by fitting equation 2 by unconditional logistic regression using data on the 142 lung cancer deaths and 714 cardiac deaths. The analysis generating column 4 is identical to that generating column 3 with the exception that the coefficient β_4 of $\ln RR(s,t)$ is no longer constrained to be 1. The estimate of β_4 is given in the row labeled " $\ln RR(s,t)$." The analysis generating column 5 is identical to that generating column 4 with the exception that the covariate $\ln RR(s,t)$ is replaced by a covariate representing a person's age in years at time of death (i.e., it represents a linear logistic model for the age effect). The estimate of the age effect is reported in the row labeled "age." Column 6 is identical to column 5 except that the covariate age^2 is added.

Column 7 represents the result of a "stratified (i.e., category-matched) analysis" in which strata are defined by joint five-year intervals of age and calendar period. The row labeled "cases" demonstrates that $142 - 139 = 3$ cases were in strata with no controls. Similarly, $710 - 665 = 45$ controls were in strata with no cases. The logistic regression model

$$\ln \left[\frac{p(D_1)}{p(D_2)} \right] = \beta_{0,k} + \beta_1 z_1 + \beta_2 z_2 + \beta_3 FB \quad (3)$$

where k is a stratum indicator was fit by

TABLE 1
Lung cancer mortality: methods of analysis of Montana copper smelter data (5)

	Full cohort		Proportionate mortality using cardiac controls					
	Cox	US rates	Unconditional				Matched	
			$\ln RR, \beta = 1$	$\ln RR$	Age	Age + age ²	5-year	10-year
Column*	1	2	3	4	5	6	7	8
Cases†	142	142	142	142	142	142	139	140
Controls‡			710	710	710	710	665	695
Covariates								
Foreign-born	0.72§ (0.20)	0.76 (0.18)	0.72 (0.21)	0.73 (0.22)	0.73 (0.23)	0.73 (0.24)	0.74 (0.22)	0.75 (0.21)
Heavy arsenic (years)	0.060 (0.013)	0.058 (0.013)	0.053 (0.017)	0.055 (0.018)	0.054 (0.017)	0.054 (0.018)	0.054 (0.019)	0.052 (0.018)
Medium arsenic (years)	0.022 (0.007)	0.022 (0.007)	0.022 (0.009)	0.022 (0.010)	0.025 (0.009)	0.025 (0.010)	0.023 (0.010)	0.022 (0.009)
$\ln RR(s,t) \ddagger$				1.15 (0.26)				
Age					-0.006 (0.007)	-0.002 (0.035)		
Age ²						-0.0006 (0.0015)		

* Each column corresponds to a different method of analysis as described in the text. The headings immediately above the column numbers are abbreviated descriptions of the analyses associated with the columns.

† The rows labeled "cases" and "controls," respectively, give the number of cases and controls used in each analysis.

‡ $\ln RR(s,t)$ is the natural logarithm of the ratio of lung cancer to cardiac mortality rates in US white males of age t in calendar year s .

§ Parameter estimates (standard errors in parentheses).

matched conditional logistic regression because of the large number of nuisance parameters, β_{0k} (12). The analysis represented in column 8 is equivalent to that represented in column 7, with the exception that strata were defined by joint 10-year levels of age and calendar period. As such, fewer cases and controls were left without matches. The entries in the body of the table give estimates of coefficients with standard errors in parentheses.

Table 2 presents results of analysis performed on a random sample of 43 lung cancer cases and 215 cardiovascular controls in order to bring out the effects of small sample sizes. To facilitate comparisons with table 1, the columns in table 2 are labeled with the column numbers of the corresponding columns in table 1.

Summary of results

We first consider table 1. In the two full cohort analyses, neither the estimates nor

standard errors of the arsenic coefficients were changed by incorporation into the analysis of the additional assumption of a constant healthy worker effect. Breslow et al. (4) show that this result is to be theoretically expected if cumulative exposure is unassociated with age and calendar period. Surprisingly, as Breslow et al. (4) show, in this data set cumulative exposure is only weakly associated with age and calendar period. The results presented in column 3 demonstrate that the effect of country of birth and of cumulative exposure to medium or heavy arsenic concentrations on mortality from cardiac disease must be small, since the estimated coefficients found in column 3 (i.e., $\beta_1, \beta_2, \beta_3$) are nearly identical to those in columns 1 and 2 (i.e., $\beta_{1,1}, \beta_{2,1}, \beta_{3,1}$). The standard errors found in column 3 are only 10 to 20 per cent greater than those in columns 1 and 2, even though data on only 842 rather than 8,042 cohort members had to be collected. This

TABLE 2
*Methods of analysis for a subsample of lung cancer cases and cardiac controls**

	Proportionate mortality					
	Unconditional				Matched	
	$\ln RR, \beta = 1$	$\ln RR$	Age	Age + age ²	5-year	10-year
Column	3	4	5	6	7	8
Cases	43	43	43	43	41	42
Controls	215	215	215	215	98	187
Covariates						
Foreign-born	0.77 (0.35)	0.79 (0.42)	0.72 (0.41)	0.73 (0.42)	0.60 (0.50)	0.69 (0.41)
Heavy arsenic (years)	0.040 (0.026)	0.040 (0.029)	0.039 (0.029)	0.039 (0.030)	0.010 (0.043)	0.033 (0.031)
Medium arsenic (years)	0.029 (0.015)	0.030 (0.018)	0.027 (0.019)	0.028 (0.028)	0.028 (0.020)	0.035 (0.018)
$\ln RR(s,t)$		1.07 (0.49)				
Age			-0.010 (0.011)	-0.005 (0.047)		
Age ²					-0.0004 (0.0025)	

* See table 1 footnotes for definition of the entries.

demonstrates the utility of case-control analyses. In the case-control analyses, only a trivial decrease in standard error is afforded by utilizing a priori information on background rates (column 3) when compared to stratifying rather finely on age and calendar period (column 7). Some theoretical work suggests that this will be the case when, as in this example, exposure is nearly uncorrelated with age and calendar period among the controls, and the exposure effect is not extreme (12, 13). Since in this data set exposure is nearly uncorrelated with age and calendar period in the controls, misspecification of the effect of age and calendar period on risk will not lead to substantial bias in the estimate of exposure effects. Thus, as expected, we see little change in the coefficients for heavy and medium arsenic exposure when modeling the age dependence of risk as linear on a logistic scale (column 5) or when stratifying

coarsely on age and calendar period (column 8).

The coefficient for the age effect is not significant in the linear logistic model represented in column 5. This reflects the fact that, although age is a strong predictor of mortality from both heart disease and lung cancer, it is only a weak predictor of the odds of dying from lung cancer versus heart disease. Column 6 shows little evidence for a nonlinear age effect. The standard errors for the exposure effects in columns 3-6 are nearly constant. In general, this can only occur if age and calendar period are neither correlates of cumulative exposure in the controls nor predictors of the mortality odds (5).

In table 2 we examine the effect of a small sample size. In column 3 of table 2, the coefficient for the effect of medium arsenic is only borderline significant ($p = 0.06$) and the coefficient for heavy arsenic

is nonsignificant ($p = 0.22$). In column 7, neither heavy nor medium arsenic are even close to being significant even though the point estimate for the effect of medium arsenic is unchanged from column 3. The increase in the standard errors in column 7 reflects the fact that over half the controls could not be matched to any case. Matching on 10-year intervals of age and calendar period markedly increased the number of controls who could be matched. Correspondingly, the standard errors of the exposure effect estimates decreased.

Even though in table 2 the effect estimates and their standard errors are seen to be relatively insensitive to the models selected for analysis (with the exception of column 7), nonetheless it is useful to have analyzed the data in several different ways. For example, if we had used only the linear logistic model for the age effect (column 5), we would have remained uncertain as to whether the estimate of the exposure effect might be biased due to failure to correctly model the possibly nonlinear dependence of the mortality odds on age, calendar period, and age-calendar period interactions. The results shown in column 3 suggest no such bias occurred.

Example 2

We now give an example in which the assumption of constant disease-specific healthy worker effects results in a marked improvement in efficiency. In this example, D_1 represents death from cardiac disease and D_2 death from all other causes. The analysis is restricted to the subcohort of copper smelter workers who were hired between 1935 and 1955 and who were age 20–22 years at time of hire. Follow-up information on this subcohort was available through 1977. In order to bring out the effect of small sample size, 30 deaths from cardiovascular disease and 60 deaths from other causes were sampled from the subcohort. In this example, we assume that the relative risk depends on “cumulative exposure” and “years since last at work” through the equation

$$\exp(\beta_{1,i}ce + \beta_{2,i}off + \beta_{3,i}off^2) \quad (4)$$

where $ce =$ (number of years of low arsenic exposure + 2 times the number of years of medium arsenic exposure + 3 times the number of years of heavy arsenic exposure); and off is the number of years since last at work.

In table 3, we use a variety of analytic approaches to estimate the parameters of equation 4 for death from heart disease. Column 1 represents the results of a nested case-control analysis within the subcohort in which, for each of the 30 cardiovascular deaths, 25 controls were sampled at random from those subcohort members who were at risk at the death age of the case and who were born within three years of the case (14, 15). For the controls, ce and off were evaluated at the death age of the case. Column 2 is similar to column 1, except that the cases represent the 60 persons dying of noncardiovascular diseases. Column 3 gives estimates of β_1 , β_2 , and β_3 (as defined following equation 2) obtained by fitting a modified version of equation 2 (in which z_1 , z_2 , FB have been replaced by ce , off , off^2) to data on the 30 cardiac cases and the 60 control deaths by unconditional logistic regression. In column 4, the coefficient of $\ln RR(s,t)$ is no longer constrained to be 1. In column 5, $\ln RR(s,t)$ is replaced by a linear age effect. In column 6, a quadratic age effect is added.

We now summarize the results found in table 3. The difference between the column 1 and column 2 estimates of the coefficients for cumulative exposure, off , and off^2 can be viewed as the estimates of β_1 , β_2 , and β_3 of the modified version of equation 2 when control for the joint effects of calendar period and age is near perfect. Columns 3–6 provide estimates of the same coefficients when various model assumptions are used to control for age and calendar period. The exposure effect estimate in column 3 is less than the difference between the column 1 and column 2 estimates by 0.014. The exposure effect estimate in column 5 exceeds this difference by 0.040. These results sug-

TABLE 3
Cardiac mortality in a subcohort of a Montana copper smelter: methods of analysis

	Nested case-control with live controls		Proportionate mortality using noncardiac controls			
	Cardiac	Other deaths	$\ln RR, \beta = 1$	$\ln RR$	Age	Age + age ²
Column	1	2	3	4	5	6
Cases	30	60	30	30	30	30
Controls	750	1,500	60	60	60	60
Covariates						
Cumulative exposure	0.023 (0.019)	0.001 (0.016)	0.008 (0.019)	0.017 (0.025)	0.062 (0.037)	0.055 (0.042)
<i>Off</i> *	0.139 (0.09)	0.082 (0.05)	0.030 (0.057)	0.044 (0.063)	0.125 (0.066)	0.109 (0.072)
<i>Off</i> ²	-0.004 (0.003)	-0.005 (0.001)	-0.001 (0.001)	-0.001 (0.001)	-0.0012 (0.001)	-0.0010 (0.001)
$\ln RR(s,t)$				0.67 (0.65)		
Age					-0.080 (0.069)	0.742 (0.413)
Age ²						-0.0080 (0.0042)

* *Off*, years since last at work. See table 1 for definition of the other entries.

gest that, as expected, the misspecification bias in the exposure effect estimate may be greater under the assumption of a linear logistic age effect than under the assumption of disease-specific constant healthy worker effects. (The residual sampling variability is too large for us to come to a definitive conclusion on this matter.) In addition, the exposure effect estimate in column 4 is closer than that in column 3 to the difference between the column 1 and column 2 effect estimates, presumably reflecting the fact that the model of column 4 incorporates fewer a priori assumptions than that of column 3. For similar reasons the effect estimate in column 6 is closer than that in column 5 to this difference. Note that there is evidence for a nonmonotone effect of the age on the mortality odds in column 6. Because the age effect in the data is nonmonotone, the linear logistic model for age (column 5) failed to detect much of an age effect.

The most striking aspect of the results

in table 3 is that the standard error for the exposure effect in column 3 is approximately 50 per cent of that in column 5 (and less than 50 per cent of that in column 6). It follows that, in order to obtain confidence intervals for the exposure effect of a given length, an investigator who assumed constant disease-specific healthy worker effects would require a sample size only one-fourth as large as that required by an investigator who assumed a linear logistic model for age. (In general, each such confidence interval would cover at its nominal rate only if the model under consideration is correctly specified.) This result reflects the strong association of cumulative exposure with age (conditional on years since last at work) in the subcohort due to the fact that all subcohort members were approximately the same age at hire. In fact, the correlation between the exposure effect estimate and the age effect estimate in column 5 is -0.85 . In typical proportionate mortality studies one would expect that the

gain in efficiency resulting from the assumption of constant healthy workers would be less extreme than the gain found in this example, but more extreme than the small gains found in our first example.

SUMMARY AND DISCUSSION

In small proportionate mortality studies if age and/or calendar period (or the interactions between age and calendar period) are strong predictors of mortality odds and/or strong correlates of the relevant exposure index in the controls, an investigator can simultaneously retain good efficiency and guard against bias by specifying realistic models for baseline mortality rates that incorporate information on baseline rates derived from US mortality tables (as in table 3, column 3). Such models can be fit to the study data using the methods shown in this paper. If age and calendar period are only weak predictors of the mortality odds and weak correlates of the relevant exposure index, then coarse stratification on age and calendar period (as in column 8, table 2) or the use of mathematically convenient parsimonious logistic models for the mortality odds (as in table 2, column 5) will efficiently control confounding.

We stress that it can be quite important to have correctly specified one's models for causes of death D_1 and D_2 with respect to all important confounders. For example, the magnitude of the healthy worker effect for heart disease measured on a ratio scale is known to decrease with increasing number of years from initial hire. To control for this effect we might, for example, add the term $\beta_4(t - t_h)$ to equation 2 where t_h is age at hire and t is age at death. Similarly, since US rates are published only for five-year intervals of age and calendar period, if $\ln RR(s, t)$ varied sharply with age and calendar period, then published US rates would probably require interpolation.

As another example, suppose data on cigarette smoking history were available for cohort members and that we could adequately model the effect of cigarette smok-

ing on lung cancer and cardiac mortality by adding the term $\beta_{4,i}PY$ to the expression in parentheses in equation 1 (where PY is pack-years). One might suppose that we could estimate the unknown parameters of equation 1 from proportionate mortality data by modifying equation 2 through the addition of the term β_4PY to the right side of the equation. The supposition is not correct since the ratio of the lung cancer and heart disease mortality rates of unexposed, native born, nonsmoking cohort members to the mortality rates of the general US white male population (which contains over 30 per cent smokers) would decrease with age. In this setting, we might consider the following modeling strategies. One strategy would be to replace the lung cancer and heart disease mortality rates of US white males with those of nonsmoking US white males (obtained from the American Cancer Society follow-up study of one million Americans) when computing the covariate $\ln RR(s, t)$ for inclusion in the modified version of equation 2 that includes the covariate β_4PY . Alternately, suppose rough estimates of the lifetime cumulative exposure to cigarettes at age t and calendar year s of an average US white male at risk at (s, t) are available from external sources. Then it might be suitable to continue to use US white male rates in computing $\ln RR(s, t)$ if, for a study subject dying at age t and calendar year s , the difference between their lifetime cumulative exposure to cigarettes and that of an average US white male alive at (s, t) was used in place of pack-years as the "smoking covariate" in the analysis.

As one last example, consider the empirical observation that unexposed persons who leave employment at any age (say, 40 years) prior to age 65 have higher age-specific mortality rates for cardiovascular disease than unexposed persons who continue at work past that age (at least, in part, because disabled workers tend to leave employment). We refer to this phenomenon as the healthy worker survivor effect. It follows that "time since last at work" is

both an independent risk factor for death and is associated with (in fact, is a determinant of) cumulative exposure (since individuals off work receive no further exposure). As such, an analysis that ignores time since last at work will tend to underestimate the effect of cumulative exposure on cardiac disease. Therefore, one may wish to adjust for time since last at work as we did in equation 4. Unfortunately, controlling for time since last at work may itself result in an underestimate of the effect of exposure on cardiac mortality, when, for certain persons, leaving employment is a proxy for the (unrecorded) onset of disabling cardiac disease. This reflects the fact that controlling for time since last at work may be tantamount to controlling for an intermediate variable (onset of disabling cardiac disease) on the causal pathway from exposure to death. If such is the case, special analytic approaches are necessary (16–18). (The above remarks would apply whether cardiac deaths are cases (as in example 2) or controls (as in example 1).)

An interesting generalization of the methods covered in this paper has recently been described by Andersen et al. (19).

Finally, we note that the methods described in this paper are easily generalized to incorporate parametric models for the background rates and generalized (i.e., nonexponential) relative risk functions. Details are provided in the Appendix. These generalizations will be essential when the baseline mortality rate for unexposed cohort members explicitly depends on the parameters of interest to the investigator. For example, in estimating the number of stages in a multi-stage model of lung cancer from proportionate mortality data, a case-control analysis that matches on age would be grossly inefficient because most of the information on the number of stages is contained in the shape of the age-incidence curve. An efficient analysis would involve modeling for lung cancer the age-exposure specific incidence curve in accordance with the predictions of the multi-stage theory.

REFERENCES

1. Miettinen OS, Wang J-D. An alternative to the proportionate mortality ratio. *Am J Epidemiol* 1981;114:144–8.
2. Prentice RL, Bollmer WM, Kalbfleisch JD. On the use of case series to identify disease risk factors. *Biometrics* 1984;40:445–58.
3. Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika* 1978;65:153–8.
4. Breslow N, Lubin JH, Marek P, et al. Multiplicative models and cohort analysis. *J Am Stat Assoc* 1983;78:1–12.
5. Robins JM. The statistical foundations of confounding in epidemiology. Technical report no. 2. Occupational Health Program. Harvard School of Public Health. October, 1983.
6. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol* 1986;123:392–402.
7. Lee AM, Fraumeni JF. Arsenic and respiratory cancer in man: an occupational study. *JNCI* 1969;42:1045–52.
8. Breslow N. The proportional hazards model: applications in epidemiology. *Commun Statist-Theor Math* 1978; A7(4):315–32.
9. Baker RJ, Nelder JA. The GLIM system: release 3. Oxford: Numerical Algorithms Group, 1978.
10. User's Guide. Cary, NC: SAS Institute, Inc, 1979.
11. Breslow N, Day N. The standardized mortality ratio. In: Sen PK, ed. *Statistics in biomedical, public health, and environmental sciences*. New York: Elsevier Science Publishing Co, 1985:55–74.
12. Breslow N, Day N. *Statistical methods in cancer research: case-control studies*. Lyon: International Agency for Research on Cancer, 1980.
13. Breslow N, Patton J. Case-control analysis of cohort studies. In: Breslow N, Whittemore A, eds. *Energy and health*. Philadelphia: Society for Industrial and Applied Mathematics, 1979:226–42.
14. Thomas DC. Appendix to Liddell FDK, McDonald JC, Thomas DC. *Methods of cohort analysis. Appraisal by application to asbestos mining (with discussion)*. *J R Stat Soc* 1977;A140:469–91.
15. Oakes D. Survival times: aspects of partial likelihood. *International Statistical Review* 1981; 49:235–64.
16. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986;7:1393–1512.
17. Robins JM. A statistical method to control for the healthy worker effect in intracohort comparisons. (Abstract.) *Am J Epidemiol* 1984;120:465.
18. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis* (in press).
19. Andersen PK, Borch-Johnsen K, Deckert T, et al. A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics* 1985;41:921–32.
20. Armitage P, Doll R. Stochastic models for carcinogenesis. In: *Proceedings 4th Berkeley Sympos-*

sium on Mathematics, Statistics, and Probability. Berkeley: University of California Press 1961:19-38.

21. Whittemore AS. The age distribution of human cancer for carcinogenic exposures of varying intensity. *Am J Epidemiol* 1977;106:418-32.
22. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403-11.

APPENDIX

In occupational epidemiology, additive excess relative risk models and multistage cancer models often have greater biologic plausibility than the exponential relative risk model of equation 1. This appendix will show how to use proportionate mortality data to estimate the parameters of these nonexponential relative risk models.

Suppose the age-specific incidence of disease *i* at age *t* in calendar year *s* can be modeled as

$$\gamma_{i,s}(t | \beta_i, Z) = r_i(\beta_i, Z) \gamma_{i,s}(t | Z = 0), \quad (A.1)$$

where β_i is a disease-specific column of regression coefficients, r_i are known (generalized relative risk) functions, Z is a vector of covariates that include an exposure of interest and relevant confounding factors and their interactions, and $Z = 0$ refers to persons with baseline levels of these covariates. Thus, $\gamma_{i,s}(t | Z = 0)$ is the mortality rate from disease *i* of an individual with baseline covariates at age *t* and calendar year *s*, and $r_i(\beta_i, Z)$ is the relative risk (i.e., the rate ratio) for an individual having covariates Z relative to someone with baseline covariates. By definition $r_i(\beta_i, Z = 0) = 1$.

Further, following Breslow et al. (4), we will assume that

$$\gamma_{i,s}(t | Z = 0) = rr_i(t, s, \theta_i) \gamma_{i,s_0}(t_0, Z = 0) \quad (A.2)$$

where rr_i are known functions, θ_i are unknown parameters, and $\gamma_{i,s_0}(t_0, Z = 0)$ is the baseline disease-specific mortality rate at some arbitrarily chosen age and calendar date (i.e., for each disease type the ratio of the age-calendar period-specific baseline mortality rates to that of the baseline mortality rate at (t_0, s_0) , $rr_i(t, s, \theta_i)$, is known except possibly for a vector of unknown parameters, θ_i). Let $\tau = (\beta_1, \beta_2, \theta_1, \theta_2)$. Then, following Prentice et al. (2), we obtain the following:

Theorem: If equation A.1 and equation A.2 hold and the data consist of random samples of deaths from causes D_1 and D_2 , then (under standard regularity conditions) asymptotically unbiased and efficient estimators of the identifiable parameters in τ can be obtained by fitting, by the method of unconditional maximum likelihood, the prospective logistic (nonlinear) regression model

$$\ln \left[\frac{p(D_1 | t, Z, s)}{p(D_2 | t, Z, s)} \right] = \alpha_0 + \ln \left[\frac{rr_1(t, s, \theta_1)}{rr_2(t, s, \theta_2)} \right] + \ln \left[\frac{r_1(\beta_1, Z)}{r_2(\beta_2, Z)} \right]. \quad (A.3)$$

Furthermore, the inverse of the formal observed or expected information of the prospective model evaluated at the parameter estimates consistently estimates the asymptotic variance of the parameter estimators.

Before sketching a proof, we show how one can use equation A.3 to fit nonlinear logistic models. To clarify our notation, we first rediscuss the exponential relative risk model used in example 1.

Example A1—an exponential relative risk model: Let $r_i(\beta_i, Z)$ be equation 1 for both disease 1 (lung cancer) and disease 2 (heart disease). Let $rr_i(t, s, \theta_i)$ be the known ratio of the mortality rate for disease *i* among US white males at age *t* and calendar year *s* to the rate among US white males at age t_0 and calendar year s_0 . (t_0 and s_0 can be chosen arbitrarily.) Then equation A.3 is exactly equivalent to equation 2 upon writing $\beta_0 = \alpha_0 - \ln RR(s_0, t_0)$ where α_0 is defined by equation A.3. (β_0 and $\ln RR(s_0, t_0)$ are as defined in and following equation 2.)

Example A2—an additive excess relative risk model: Example A2 is similar to example A1, except that we shall replace the exponential relative risk model with the additive excess relative risk model $r_i(\beta_i, z) = 1 + \beta_{i,z}$ where *z* is cumulative exposure to arsenic. Then, the final term on the right side of equation A.3 is of the form

$$\ln[(1 + \beta_{1,z}) / (1 + \beta_{2,z})] \quad (A.4)$$

rather than

$$\ln[\exp\{(\beta_{1,z} - \beta_{2,z})z\}] \quad (A.5)$$

as in an exponential relative risk model. In an exponential relative risk model, only the function in equation A.5 can be estimated (i.e., is identifiable). Thus, $\beta_{1,z}$ and $\beta_{2,z}$ are not separately identifiable, since the function in equation A.5 depends only on their difference. In contrast, in the additive excess relative risk model, $\beta_{1,z}$ and $\beta_{2,z}$ are both identifiable, provided $\beta_{1,z} \neq \beta_{2,z}$, since knowledge of the estimable function in equation A.4 allows one to compute both $\beta_{1,z}$ and $\beta_{2,z}$. Thus, in theory, if one knew that the dose-response for causes of death D_1 and D_2 was of an additive excess relative risk form, one could estimate the effect of exposure on cause of death D_1 (i.e., estimate $\beta_{1,z}$) from proportionate mortality data, even when exposure was a risk factor for cause of death D_2 (i.e., $\beta_{2,z} \neq 0$)! Unfortunately, if one even slightly misspecified the form of the dose-response curve (as is inevitable), one's inferences concerning the effect of exposure on D_1 would be seriously in error if exposure was a strong risk factor for D_2 . Thus, in practice, we suggest *never* using, as a control disease, a disease for which exposure is believed to be a risk factor. It follows that in actually fitting an additive excess relative risk model we would use $\ln(1 + \beta_{1,z})$ as the final term on the right side of equation A.3. The SAS procedure NLIN can be used to carry out the computations.

Example A3: Suppose one assumes a priori that for cause of death D_1 , lung cancer, exposure affected a single stage of a *k* stage multi-stage cancer process as

originally described by Armitage and Doll (20). Then, $r_1(\beta_1, Z)rr_1(t, s, \theta_1) = t^{k-1}/t_0^{k-1} + \alpha_2 \int_0^t d(u)(t-u)^{k-i-1}u^{i-1}du$ where $d(u)$ is the dose at time u , i is the stage affected, and α_2 is an unknown constant (21). Suppose that the control disease, D_2 , is heart disease and that we assume $r_2(\beta_2, z) = 1$ and $rr_2(t, s, \theta_2)$ is known and is as in example A1. Then the unknown parameters $\tau = (i, k, \alpha_2)$ can be estimated, using the SAS procedure NLIN, from the nonlinear logistic regression model of equation A.3 upon recognizing that we can rewrite the right side of equation A.3 as $\alpha_0 + \ln(r_1(\beta_1, z)rr_1(t, s, \theta_1)) - \ln(r_2(\beta_2, z)rr_2(t, s, \theta_2))$.

Finally, we sketch a heuristic proof of the theorem. Let $x = (t, s, Z)$ represent age, calendar date, and covariate history up to (s, t) . Define $x_0 = (t_0, s_0, Z = 0)$ and $\Delta_i = (\beta_i, \theta_i)$. Let $f_i(x, \Delta_i) = r_i(\beta_i, Z)rr_i(t, s, \theta_i)$ so that $f_i(x_0, \Delta_i) = 1$. We view the observed cohort as having been sampled from a large near-infinite superpopulation. Let g_j be the total number of years individual j was at risk (i.e., under active follow-up). Independent left, right, and interval censoring are allowed but do not contribute to g_j . Let $G = \sum g_j$ where the sum is over all members of the superpopulation. Consider the following two-stage procedure for selecting a value of x . Sample a person at random from the superpopulation such that the probability of selecting individual j is g_j/G . Next, for the sampled individual j , choose g at random from the uniform distribution on $[0, g_j]$. Let x be the value of x for individual j after g years at risk. x has a well defined distribution under the above two-stage procedure with density $h(x)$. Then, equation A.1 and A.2 imply

$$p(x | D_i) = \frac{f_i(x, \Delta_i)h(x)}{\int f_i(x, \Delta_i)h(x) dx} \tag{A.6}$$

Therefore,

$$\frac{p(x | D_1)/p(x_0 | D_1)}{p(x | D_2)/p(x_0 | D_2)} = \frac{f_1(x, \Delta_1)}{f_2(x, \Delta_2)} = \gamma(x, \Delta_1, \Delta_2) \tag{A.7}$$

But A.7 is exactly equivalent to equations 3 and 5 of Prentice and Pyke (22), except that the exponential relative risk function is replaced by the generalized relative risk function $\gamma(x, \Delta_1, \Delta_2)$. But the results of Prentice and Pyke are valid for any generalized relative risk function. In addition, their results only required that their equations 3 and 5 held (our A.7 in the generalized relative risk context) and the statistical independence of the observations. Thus, our theorem follows immediately from the results of Prentice and Pyke.

The above proof suggests that, if full cohort data were available, we could use the following case-control design: 1) Sample controls j at random from the observed cohort with control selection probabilities g_j/G where now the sum defining G is over the subjects in the observed cohort. 2) Choose a value of x for each sampled control j as in the above proof of the theorem. 3) Using the SAS procedure NLIN, fit equation A-3, modified so that D_2 refers to the controls chosen above, $rr_2(t, s, \theta_2) = 1$, and $r_2(\beta_2, Z) = 1$.

Evaluation of the statistical properties of this design would be of interest.