American Journal of Epidemiology

Vol. 188, No. 2
DOI: 10.1093/aje/kwy223
Advance Access publication:
October 19, 2018

Practice of Epidemiology

General Relative Rate Models for the Analysis of Studies Using Case-Cohort Designs

David B. Richardson*, Bryan Langholz, and Kaitlin Kelly-Reif

* Correspondence to Dr. David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (e-mail: david.richardson@unc.edu).

Initially submitted January 4, 2018; accepted for publication September 25, 2018.

A standard approach to analysis of case-cohort data involves fitting log-linear models. In this paper, we describe how standard statistical software can be used to fit a broad class of general relative rate models to case-cohort data and derive confidence intervals. We focus on a case-cohort design in which a roster has been assembled and events ascertained but additional information needs to be collected on explanatory variables. The additional information is ascertained just for persons who experience the event of interest and for a sample of the cohort members enumerated at study entry. One appeal of such a case-cohort design is that this sample of the cohort may be used to support analyses of several outcomes. The ability to fit general relative rate models to case-cohort data may allow an investigator to reduce model misspecification in exposure-response analyses, fit models in which some factors have effects that are additive and others multiplicative, and facilitate estimation of relative excess risk due to interaction. We address model fitting for simple random sampling study designs as well as stratified designs. Data on lung cancer among radon-exposed men (Colorado Plateau uranium miners, 1950–1990) are used to illustrate these methods.

cohort studies; Cox model; dose-response function; models, statistical; radiation; software, statistical

Abbreviations: CI, confidence interval; WLM, working level month.

A case-cohort study design may be employed as an approach to analysis of exposure-disease associations in epidemiologic studies where a roster of cohort members has been assembled and followed to ascertain events but additional information needs to be collected on at least 1 explanatory variable (1). Ascertaining information about explanatory variables in a large cohort may be expensive or time-intensive. Rather than obtaining that additional information for the full cohort, information can be ascertained just for persons who experience the event of interest and for a sample of persons drawn from the cohort roster at baseline (i.e., a subcohort). The case-cohort design has been used in a wide range of epidemiologic studies (2–6), because it may facilitate collection of information on an exposure or covariate of interest (1, 7) and because it allows for the possibility that the same subcohort may be used for analyses of several outcomes (8, 9). The latter is a feature of the case-cohort design that distinguishes it from a nested case-control study, in which the control series is matched to an index disease.

Previous authors have described how standard statistical software packages can be used to fit the Cox proportional hazards regression model to case-cohort data (10–12). The procedures available in most standard statistical packages, such as SAS (SAS Institute, Inc., Cary, North Carolina) and Stata (StataCorp LLC, College Station, Texas), for fitting Cox proportional hazards regression models limit a data analyst to log-linear models of the form $\varphi(\boldsymbol{Z};\boldsymbol{\beta}) = \exp(\boldsymbol{Z\beta})$, where $\boldsymbol{Z}$ is a vector of explanatory variables and φ is the hazard ratio. This implies exponential exposure-response trends and multiplicative interactions. In a review of the literature, Cologne et al. (13) noted that only 1 specialized software package is currently available to fit general relative rate models for case-cohort data that are not constrained to the log-linear model form, and users are restricted to fitting such models to data derived from a simple random sampling case-cohort design. Here we describe how standard statistical software packages may be used to fit this broader class of general relative rate models. Model forms other than the log-linear model are desirable when they provide a better representation of the exposure response or when they address biological or public health questions (14, 15). They also may facilitate exposure-response analyses, since misspecification of model form can lead to loss of

statistical power for a model-based test of an exposure-response association and to the possibility that estimates of effect for extreme exposure levels will be substantially distorted. For example, a linear relative rate model, under which the rate of disease increases or decreases in an additive fashion with exposure, may be of interest in analyses of environmental or occupational exposure (16, 17). In this paper, we illustrate how general relative rate models may be readily fitted to case-cohort data using the SAS statistical software package.

METHODS

Consider a cohort study in which n people are followed over time to ascertain a binary outcome of interest. Let A and T denote study entry and exit times, respectively, let Y denote the binary outcome, and let vector $\boldsymbol{Z}$ denote explanatory variables of interest, which can be time-dependent.

Suppose that we sample m members from the full cohort to form a subcohort by simple random sampling. Let S denote a binary variable with probability $S = 1$ equal to m/n; if S takes a value of 1, then the person is a member of the subcohort. The case-cohort analysis will include members of the subcohort ($S = 1$), for whom information on $\boldsymbol{Z}$ is collected. In addition, the case-cohort analysis will include any person who experiences the outcome but is not in the subcohort (i.e., people for whom $Y = 1$ and $S = 0$), for whom information on $\boldsymbol{Z}$ is collected as well. A data structure for this information may include 1 record per person in the case-cohort study, indicating, for each study member, A, T, $\boldsymbol{Z}$, Y, and S.

Following the approach in prior work in which we focused on fitting general relative rate models to cohort and nested case-control data (18), we proceed by assuming that the underlying population model for the hazard rate of the outcome that generates the observed data conforms to the form

$$\lambda(t, \boldsymbol{Z}(t)) = \lambda_0(t)\varphi(\boldsymbol{Z}(t); \boldsymbol{\beta}),$$

where $\lambda(t, \boldsymbol{Z}(t))$ denotes the (instantaneous) hazard rate at time t among subjects with covariate vector $\boldsymbol{Z}(t)$, $\lambda_0(t)$ denotes the time-dependent baseline hazard rate, and $\boldsymbol{\beta}$ denotes a vector of population parameters for the relative hazard rate that we wish to estimate. The function $\varphi(\cdot)$ can encompass a wide range of models other than the form $\varphi(\boldsymbol{Z}(t); \boldsymbol{\beta}) = \exp(\boldsymbol{Z}(t)\boldsymbol{\beta})$, which is commonly used. The linear excess relative rate model, $\varphi(\boldsymbol{Z}(t); \boldsymbol{\beta}) = 1 + \boldsymbol{Z}(t)\boldsymbol{\beta}$, is one model form of interest, particularly in environmental and occupational epidemiology. More generally, prior authors have described a class of relative risk models that permit a data analyst to compare linear and log-linear dose-response models or additive and multiplicative interaction models by incorporating each as a special case of a broader set of models under consideration, such as a model of the form $\varphi(\boldsymbol{Z}(t); \boldsymbol{\beta}) = [\exp(\boldsymbol{Z}(t)\boldsymbol{\beta})]^{1-\theta}[1 + \boldsymbol{Z}(t)\boldsymbol{\beta}]^{\theta}$, with θ a parameter that equals 1 if the model is completely additive and 0 if the model is multiplicative (16, 19, 20).

The standard Cox partial likelihood, which is used for the full cohort design, must be adjusted for application to the case-cohort design. Various weighting schemes have been proposed for this adjustment (21). In this paper, we use the weighting scheme proposed by Prentice (1), which applies a weight of 1 to all subjects at all times except for nonsubcohort cases, who are assigned a weight of 0 at all times prior to their failure.

Implementation

We follow the approach described by Langholz and Jiao (11). Briefly, the weights proposed by Prentice (1) result in a risk set formed at each failure time that includes the case and all those subcohort members in the study and at risk at the failure time, while the cases outside of the subcohort enter the risk sets only at their event time. Rather than using a weighting scheme to implement this adjustment to the risk sets, we use an equivalent approach, which is to adjust the entry times for nonsubcohort cases (11).

First, we adjust the entry times for cases outside of the subcohort (i.e., nonsubcohort cases) so that these cases enter the study just prior to their exit times. This ensures that a nonsubcohort case (i.e., $Y = 1$ and $S = 0$) does not appear in the risk sets for prior event failure times; consequently, the nonsubcohort case does not contribute information to any risk set other than the one enumerated for its own failure time. Second, risk sets are enumerated on the basis of the adjusted entry times and observed failure times. Third, a data structure is generated with 1 record per risk set (i.e., 1 record for each case and its affiliated risk set members). We previously described in detail how such a data structure can be generated (18). The covariate information is represented by the variables z_1, z_2, …, z_k, where z_1 denotes the case's covariate information and $z_2 \ldots z_k$ denote the other risk set members' covariate information. Table 1 illustrates the case-cohort data structure for a hypothetical case-cohort study with 2 cases, one of which is a nonsubcohort case. Table 2 illustrates the risk sets constructed on the basis of the adjusted entry times and observed failure times, and Table 3 depicts the final analytical data structure generated with 1 record per risk set.

Table 1. A Case-Cohort Data Structure for a Hypothetical Case-Cohort Study With 2 Cases[a]

ID	Variable				
	A	*T*	*S*	*Y*	*Z*
1	38.4	86.6	1	0	0.8
2	35.0	81.1	1	0	1.0
3	29.3	77.6	1	0	1.1
4	31.8	77.4	1	0	0.0
5	27.8	78.5	1	0	1.3
6	53.9	90.0	1	0	0.0
7	42.5	87.1	1	0	0.0
8	25.6	77.1	1	0	0.0
9	43.0	77.9	0	1	0.0
10	37.7	76.8	1	1	1.6

Abbreviation: ID, identifier.

[a] "ID" is a unique identifier for each study participant; *A* and *T* denote study entry and exit times, respectively; *S* is a binary variable that takes the value 1 if the person is a member of the subcohort and 0 otherwise; *Y* is the binary outcome variable; and *Z* is an explanatory variable of interest.

Table 2. A Long–Risk-Set Data Structure for a Hypothetical Case-Cohort Study With 2 Cases[a]

SETNO	Variable		
	ID	Y	Z
1	10	1	1.6
1	8	0	0.0
1	4	0	0.0
1	3	0	1.1
1	5	0	1.3
1	2	0	1.0
1	1	0	0.8
1	7	0	0.0
1	6	0	0.0
2	9	1	0.0
2	5	0	1.3
2	2	0	1.0
2	1	0	0.8
2	7	0	0.0
2	6	0	0.0

Abbreviations: ID, identifier; SETNO, set number.

[a] "SETNO" indexes the risk set formed at each failure time; "ID" is a unique identifier for each study participant; *Y* is the binary outcome variable; and *Z* is an explanatory variable of interest.

The analytical data structure shown in Table 3 can be easily generated from the case-cohort data. SAS code, provided in Web Appendix 1 (available at https://academic.oup.com/aje), creates the necessary data structure illustrated in Table 3, including a macrovariable "&maxsetsize" that indicates the maximum size of any risk set in the study, a variable "ntot" that indicates the size of each risk set (i.e., the total number of people in the risk set), and a variable "y" which serves as the outcome variable for regressions.

Estimation of the parameters for the function $\varphi(\cdot)$ for case-cohort data can be performed by identifying the value of $\hat{\beta}$ that maximizes the pseudolikelihood, where, assuming no tied failure times, the contribution to the risk set for case i who fails at time t, given our case-cohort study data, is

Table 3. A Wide Data Structure With 1 Record per Risk Set for a Hypothetical Case-Cohort Study With 2 Cases[a]

SETNO	Variable										
	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	n_{tot}	y
1	1.6	0.0	0.0	1.1	1.3	1.0	0.8	0.0	0.0	9	1
2	0.0	1.3	1.0	0.8	0.0	0.0				6	1

Abbreviation: SETNO, set number.

[a] "SETNO" indexes the risk set formed at each failure time, and $z_1 \ldots z_9$ denote covariate information, where z_1 denotes the case's covariate information and $z_2 \ldots z_9$ denote the other risk set members' covariate information. The variable n_{tot} indicates the size of each risk set (i.e., the total number of people in the risk set), and y is set to 1 and serves as the outcome variable associated with each record.

$$L(\beta) = \frac{\varphi_i(\beta)}{\sum_{k \in \tilde{R}_i} \varphi_k(\beta)},$$

where $\tilde{R}_i$ denotes the "sampled risk set" for case i at failure time t_i given a subcohort of size m (i.e., restricting to people for whom the baseline variable $S = 1$), and $\varphi_k(\beta) = \varphi(Z_k(t); \beta)$. The SAS NLMIXED procedure is used to obtain maximum likelihood estimates. For example, a regression analysis of the data, once organized as in Table 3, employing the linear excess relative rate model form could be fitted via SAS as follows:

```
proc nlmixed data= ;
  parms beta=0 ;
   array z{&maxsetsize}
     z1-z&maxsetsize;
   sum=0;
  do i = ntot to 1 by -1;
   Phi= (1 + z{i}*beta );
   sum = sum + Phi ;
  end;
  L= Phi/sum;
  model y ~ general(log(L)); run;
```

Here, the "parms" statement sets an initial value for the parameter "beta" to be estimated. The array "z{}" indexes the covariate values. In this example, "phi" conforms to a linear excess relative rate model form. Other model forms can be fitted easily; for example, the log-linear model would be fitted simply by replacing the statement "Phi = 1 + z{i}*beta;" with the statement "Phi = exp(z{i}*beta);".

Variance estimation under the case-cohort design is an important consideration because the standard errors for the estimated coefficients outputted by the SAS code shown above fail to account for the outcome-dependent sampling in the case-cohort data. A robust variance estimator for regression estimates under the case-cohort setting was proposed by Barlow et al. (10) of the form

$$V(\hat{\boldsymbol{\beta}}) = \sum_j \Delta\hat{\boldsymbol{\beta}}_j \Delta\hat{\boldsymbol{\beta}}_j^T,$$

where $\Delta\hat{\boldsymbol{\beta}}_j$ denotes the estimated change in $\hat{\boldsymbol{\beta}}$ when the jth individual is deleted. This quantity is often referred to as a dfbeta residual. The variance estimator corresponds to a sum of squared changes in the parameter estimate vector when observations are removed from the analysis one at a time. Following the approach of Langholz and Jiao (11), robust variance estimates can be obtained as these sums of squares (and cross-products) of the dfbeta residuals for subjects in the case-cohort study.

Web Appendix 2 provides SAS code with which to obtain the robust variance estimator. We use the "predict" command in PROC NLMIXED to predict "phi" and its derivative for each person (i.e., using a separate predict statement for each person) and output these to a file. To compute the robust variance estimator, first dfbeta values for all subjects (based on study identifier (id)) are calculated, and then the dfbeta values are summed using the PROC SUMMARY procedure. The robust variance estimator

is computed using PROC CORR, requesting that the sums of squares and cross-products be outputted (sscp option). Given a robust variance estimator, a Wald-type 95% confidence interval can be computed in the usual way.

Prior authors have noted that the distribution of the maximum likelihood estimators for non–log-linear models, such as the linear excess relative rate model, may be far from normal unless the sample size is large (17, 21). In such settings, a Wald-type 95% confidence interval may not perform as desired. For non–log-linear models, the Wald-type 95% confidence interval may not have nominal coverage. Barlow (19) and Prentice and Mason (17) suggested parameter transformations to improve asymptotic distributional approximations for the linear excess relative rate model.

Bootstrapping is another method that can be employed to derive confidence intervals for the estimated parameters, with percentile-based bootstrap confidence intervals obtained without assumptions about the distributional form of the maximum likelihood estimators (22, 23). We adapt the weighted bootstrap, which assigns a nonnegative random weight to each person; following the method of Huang (22), and similar to the Bayesian bootstrap approach of Rubin (24), we assign weights by random sampling from the standard exponential distribution such that these weights are independent of the data with unit mean and unit variance. This approach creates new data sets through reweighting of the initial data. To fit the weighted regression models, we use the following weighted likelihood:

$$L(\beta) = \frac{[\varphi_i(\beta)]^{w_i}}{[\sum_{k\in\tilde{R}_i} \varphi_k(\beta) w_k]^{w_i}},$$

where w_i denotes weights that are assigned by random sampling for study member i.

The weighted bootstrap proceeds by assigning B random weights to each person. To obtain a bootstrap estimator, we fit B weighted regression models on the new data sets generated through reweighting of the initial data. Because these weights are drawn from the exponential distribution, no person is assigned a weight equal to, or less than, 0; therefore, all of the case failures in the full cohort are still observed for each bootstrap sample, and the number of risk sets does not change, nor do the observed failure times. Consequently, the case-cohort risk sets only need to be enumerated once, after which a large number of weighted regressions may be fitted. A Wald-type confidence interval may be derived by taking an estimate of the variance of the bootstrap estimates as an estimator of the variance of the estimated parameter. Alternatively, a percentile-based bootstrap confidence interval may be obtained from appropriate quantiles of the distribution of the estimates for the parameter of interest; for example, we can derive the 95% bootstrap confidence interval as the 2.5% and 97.5% quantiles of the distribution of the estimates of the parameter of interest. Web Appendix 3 provides SAS code for bootstrapping confidence intervals.

Case-cohort designs that are stratified on a covariate

Suppose that the subcohort has been sampled within defined strata and that the investigator wishes to fit a stratified proportional hazards regression model with separate stratum-specific baseline hazards. Following the approach of Langholz and Jiao (11), we will refer to this as a confounder-stratified case-cohort design, noting that this stratified case-cohort design has also been studied by Borgan et al. (25) and Samuelsen et al. (26). A stratified random sample implies that the subcohort is selected such that the sampling probability depends upon covariate strata, c. To illustrate, suppose n_c and m_c denote the total number of subjects and the number of subjects sampled from covariate stratum c, respectively. Let S_c denote a binary variable with probability $S_c = 1$ equal to m_c/n_c; if S_c takes a value of 1, then the person is a member of the subcohort in stratum c. Analysis proceeds with each stratum contributing independent pseudolikelihood factors to the overall pseudolikelihood, with each factor being computed on the case-cohort members from the stratum. Thus, analysis may be conducted using the same SAS code as that presented above; and robust (11) and bootstrap variance estimates for a confounder-stratified case-cohort design may be derived using the same approach as that described for a simple random sample case-cohort design. Web Appendix 4 provides SAS code for bootstrapping confidence intervals for a stratified case-cohort design.

Empirical example

As an empirical example, we use data from a study of lung cancer among 2,704 men employed in underground uranium mining operations on the Colorado Plateau between January 1, 1950, and December 31, 1960 (27, 28). Vital status was ascertained through December 31, 1990. The outcome of interest, lung cancer mortality, was defined by underlying cause of death. The primary exposure of interest was defined as cumulative radon exposure, lagged 5 years, expressed in working level months (WLMs) and computed for each worker as the product of the duration of employment in each job in a year and the estimated rate of radon exposure for that job (29). Miners who began working prior to 1950 were not included, since radon measurements were not available for that period and therefore radon exposure could not be reliably estimated during that time. The cohort included 263 lung cancer cases; the cumulative radon exposure accrued among lung cancer cases ranged from 2 WLMs to 20,870 WLMs, and the mean and median cumulative exposure among cases were 1,292 WLMs and 791 WLMs, respectively. The mean and median cumulative exposure among noncases were 586 WLMs and 309 WLMs, respectively.

The example involves a case-cohort study drawn from this cohort. The case-cohort sample included a subcohort of 500 miners randomly sampled from the full cohort. The subcohort included 48 lung cancer cases and 452 noncases; the case-cohort data included the subcohort plus the 215 lung cancer cases that were not included in the subcohort. First, we fitted a standard log-linear proportional hazards model for the association between cumulative radon exposure (lagged 5 years) and lung cancer of the form $\varphi(Z(t);\beta) = \exp(Z(t)\beta)$; we obtained a point estimate and associated robust 95% confidence interval using SAS PROC PHREG with a robust variance estimator (invoked by the "covsandwich" option) implemented using the approach for case-cohort analysis described by Langholz and Jiao (11). We also used the approach described in this paper to obtain a point estimate for the parameter β

Table 4. Estimated Association Between Radon Exposure and Lung Cancer Mortality Obtained From Fitting a Log-Linear Regression Model to Data on Colorado Plateau Uranium Miners, 1950–1990[a]

Model	$\hat{\beta}$	95% CI Estimate			
		Robust 95% CI[b]	Robust 95% CI[c]	Bootstrap 95% CI[d]	Bootstrap 95% CI[e]
$\varphi(Z;\beta) = \exp(\beta Z)$	0.04	0.03, 0.05	0.03, 0.05	0.03, 0.05	0.03, 0.05

Abbreviation: CI, confidence interval.
[a] We used data from a study of lung cancer among 2,704 men employed in underground uranium mining operations on the Colorado Plateau between January 1, 1950, and December 31, 1960 (27, 28).
[b] Robust 95% CI obtained using SAS PHREG with the "covsandwich" option as in the article by Langholz and Jiao (11).
[c] Robust 95% CI obtained using the approach described in this paper.
[d] Bootstrapped 95% CI based on the bootstrap variance and Wald-type interval.
[e] Bootstrapped 95% CI based on percentiles of the bootstrap estimates.

and associated robust and bootstrap 95% confidence intervals (based on 500 bootstrap samples).

Next, to illustrate a model form that cannot be fitted using standard statistical programs for log-linear regression models, we fitted a linear excess relative hazards model for the association between cumulative radon exposure and lung cancer of the form $\varphi(Z(t);\beta) = 1 + \beta Z(t)$. The estimated value, $\hat{\beta}$, describes the excess relative rate of lung cancer per unit of exposure; we use the approach described in this paper to obtain associated robust and bootstrap 95% confidence intervals for the estimated excess relative rate. In addition, Barlow (19) and Prentice and Mason (17) suggested parameter transformations to improve asymptotic distributional approximations for the linear excess relative rate model. Given a single regressor variable, $\boldsymbol{Z}(t) = Z(t)$, with support $[0, x_0)$, where $0 < x_0 < \infty$, the possible range of values for β under a linear excess relative rate model is $(-x_0^{-1}, \infty)$. The transformation $\alpha = \log(\beta + \beta_0)$, or equivalently $\beta = (e^{\alpha} - \beta_0)$, removes this range restriction on β; in the proposed transformation, β_0 equals x_0^{-1}, where the value of x_0 is known in a given study because it corresponds to the maximum value of the regressor variable of interest. We fitted a model of the form $\varphi(Z(t);\beta) = 1 + (e^{\alpha} - \beta_0)Z(t)$, which we expected to bring about some improvement in asymptotic distributional approximations allowing construction of the Wald-type interval (17). The estimated value, $e^{\hat{\alpha}} - \beta_0$, describes the excess relative rate of lung cancer per unit of exposure, and we constructed Wald-type confidence intervals using the standard error (SE) as $e^{\hat{\alpha}\pm 1.96\widehat{\mathrm{SE}}(\hat{\alpha})} - \beta_0$.

RESULTS

First we estimated the association between radon exposure and lung cancer using a log-linear proportional hazards regression model with a robust variance estimator. We obtained an estimate of the change in the log hazard rate ratio per 100 WLMs (log[hazard rate ratio per 100 WLMs] = 0.04) and its associated robust 95% confidence interval (95% confidence interval (CI): 0.03, 0.05). We also used the approach described in this paper to fit a log-linear proportional hazards model, which yielded an estimate of the change in the log hazard rate ratio per 100 WLMs (log[hazard rate ratio per 100 WLMs] = 0.04), a robust 95% confidence interval (95% CI: 0.03, 0.05), a bootstrap 95% confidence interval using the bootstrap variance to derive a Wald-type interval (95% CI: 0.03, 0.05), and a bootstrap 95% confidence interval based on percentiles of the bootstrap estimates (95% CI: 0.03, 0.05). For the log-linear model, the estimated coefficients and confidence intervals from a standard method and the proposed method were identical (Table 4).

Next we fitted a linear excess relative rate regression model using our proposed approach (Table 5). This yielded an estimated excess relative rate of 0.33 per 100 WLMs; the robust 95%

Table 5. Estimated Association Between Radon Exposure and Lung Cancer Mortality Obtained From Fitting Excess Relative Rate Regression Models to Data on Colorado Plateau Uranium Miners, 1950–1990[a]

Model	ERR/100 WLMs	95% CI Estimate		
		Robust 95% CI[b]	Bootstrap 95% CI[c]	Bootstrap 95% CI[d]
$\varphi(Z(t);\beta) = 1 + \beta Z(t)$	0.33	0.06, 0.59	0.04, 0.77	0.15, 0.82
$\varphi(Z(t);\beta) = 1 + (e^{\alpha} - \beta_0)Z(t)$	0.33	0.15, 0.73	0.15, 0.81	0.15, 0.82

Abbreviations: CI, confidence interval; ERR, excess relative rate; WLM, working level month.
[a] We used data from a study of lung cancer among 2,704 men employed in underground uranium mining operations on the Colorado Plateau between January 1, 1950, and December 31, 1960 (27, 28).
[b] Robust 95% CI obtained using the approach described in this paper.
[c] Bootstrapped 95% CI based on the bootstrap variance and Wald-type interval.
[d] Bootstrapped 95% CI based on percentiles of the bootstrap estimates.

confidence interval (95% CI: 0.06, 0.59) differed somewhat from the bootstrap confidence interval obtained using the bootstrap variance to derive a Wald-type interval (95% CI: 0.04, 0.77) and differed more markedly from a bootstrap confidence interval based on percentiles of the distribution of the bootstrap estimates (95% CI: 0.15, 0.82).

Finally, we estimated the excess relative rate per 100 WLMs by fitting a model with a transformation of the parameter of interest, as suggested by Prentice and Mason (17) (Table 5). This also yielded an estimated excess relative rate of 0.33 per 100 WLMs; the robust 95% confidence interval was 0.15, 0.73, and the bootstrap confidence interval obtained using the bootstrap variance to derive a Wald-type interval (95% CI: 0.15, 0.81) was very similar to the bootstrap confidence interval based on percentiles of the distribution of the bootstrap estimates (95% CI: 0.15, 0.82).

DISCUSSION

A standard approach to analysis of case-cohort data under the proportional hazards model is to fit a log-linear regression model and obtain robust confidence intervals for the estimated parameters. Prior authors have described how analysis can proceed using standard statistical packages to fit log-linear models and obtain robust confidence intervals (10–12). There may be interest in approaches to fitting alternatives to log-linear proportional hazards models to case-cohort data. For example, in epidemiologic studies of exposures to asbestos, benzene, radon progeny, and external ionizing radiation, investigators have found support for modeling the excess relative rate per unit of exposure as a linear function of exposure rather than an exponential function of exposure (2–5). One obstacle to fitting general relative hazards models to case-cohort data has been implementation using standard statistical packages (13). Analysts have tended to use specialized software that was written specifically for fitting models of this form to epidemiologic data. This paper describes an approach for fitting a broader class of models to case-cohort data that have been described as general relative rate models (16, 19, 20). We have used SAS software to illustrate the methods, but we hope that they will be implemented using other software packages as well.

For maximum pseudolikelihood estimation of parameters for general relative rate models using case-cohort data, we have described how robust and bootstrap confidence intervals may be obtained. As prior authors have pointed out, when the relative risk is not multiplicative, the distribution of the maximum likelihood estimators may be far from normal unless the sample size is large (21). Prior work has shown that the bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality (30) and that parameter transformations can somewhat improve asymptotic distributional approximations for the linear excess relative rate model (21).

In our example, for the log-linear models, Wald-type confidence intervals based on robust variance estimators were essentially identical to bootstrap confidence intervals obtained using the bootstrap variance to derive a Wald-type interval and to the bootstrap confidence interval based on percentiles of the distribution of the estimates (Table 4). For the linear excess relative model, which is a non–log-linear model form, Wald-type confidence intervals based on robust variance estimators and Wald-type intervals obtained using the bootstrap variance differed somewhat from a percentile-based bootstrap confidence interval. This was anticipated because for non–log-linear models the distribution of the maximum likelihood estimators may be far from normal unless the sample size is large. When we used a parameter transformation suggested by Prentice and Mason (17) to improve asymptotic distributional approximations for the estimates obtained under the linear excess relative rate model, Wald-type confidence intervals based on robust variance estimators were similar to bootstrap confidence intervals obtained using the bootstrap variance to derive a Wald-type interval and were very similar to the bootstrap confidence interval based on percentiles of the distribution of the estimates (Table 5). Given the relative ease of implementing the bootstrap approach described here, a percentile-based bootstrap confidence interval may be desirable when fitting non–log-linear regression models to case-cohort data.

This paper facilitates fitting of general relative rate models to case-cohort data. We have described how such models may be fitted using a standard statistical package, addressed confounder-stratified case-cohort designs, and described how to obtain confidence intervals for estimated parameters.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (David B. Richardson, Kaitlin Kelly-Reif); and Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California (Bryan Langholz).

D.B.R. was supported in part by National Institute of Occupational Safety and Health grant R03 OH010946.

Conflict of interest: none declared.

REFERENCES

1. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73(1):1–11.
2. Petersen GR, Gilbert ES, Buchanan JA, et al. A case-cohort study of lung cancer, ionizing radiation, and tobacco smoking among males at the Hanford Site. *Health Phys*. 1990;58(1):3–11.
3. Rericha V, Kulich M, Rericha R, et al. Incidence of leukemia, lymphoma, and multiple myeloma in Czech uranium miners: a case-cohort study. *Environ Health Perspect*. 2006;114(6):818–822.
4. Stenehjem JS, Kjærheim K, Bråtveit M, et al. Benzene exposure and risk of lymphohaematopoietic cancers in 25,000 offshore oil industry workers. *Br J Cancer*. 2015;113(11):1641.
5. Checkoway H, Lundin JI, Costello S, et al. Possible procarcinogenic association of endotoxin on lung cancer among Shanghai women textile workers. *Br J Cancer*. 2014;111(3):603–607.

6. Sharp SJ, Poulaliou M, Thompson SG, et al. A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS One*. 2014;9(6):e101176.
7. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc*. 1975;70(351):524–528.
8. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*. 1991;2(2):155–158.
9. Kulathinal S, Karvanen J, Saarela O, et al. Case-cohort design in practice—experiences from the MORGAM Project. *Epidemiol Perspect Innov*. 2007;4:15.
10. Barlow WE, Ichikawa L, Rosner D, et al. Analysis of case-cohort designs. *J Clin Epidemiol*. 1999;52(12):1165–1172.
11. Langholz B, Jiao J. Computational methods for case-cohort studies. *Comput Stat Data Anal*. 2007;51(8):3737–3748.
12. Therneau TM, Li H. Computing the Cox model for case cohort designs. *Lifetime Data Anal*. 1999;5(2):99–112.
13. Cologne J, Preston DL, Imai K, et al. Conventional case-cohort design and analysis for studies of interaction. *Int J Epidemiol*. 2012;41(4):1174–1186.
14. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol*. 2009;169(6):756–760.
15. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology*. 2009;20(6):863–871.
16. Thomas D. General relative risk models for survival time and matched case-control analysis. *Biometrics*. 1981;37(4):673–686.
17. Prentice RL, Mason MW. On the application of linear relative risk regression models. *Biometrics*. 1986;42(1):109–120.
18. Langholz B, Richardson DB. Fitting general relative risk models for survival time and matched case-control analysis. *Am J Epidemiol*. 2010;171(3):377–383.
19. Barlow WE. General relative risk models in stratified epidemiologic studies. *Appl Stat*. 1985;34(3):246–257.
20. Breslow NE, Storer BE. General relative risk functions for case-control studies. *Am J Epidemiol*. 1985;122(1):149–162.
21. Moolgavkar SH, Venzon DJ. General relative risk regression models for epidemiologic studies. *Am J Epidemiol*. 1987;126(5):949–961.
22. Huang Y. Bootstrap for the case-cohort design. *Biometrika*. 2014;101(2):465–476.
23. Wacholder S, Gail M, Pee D, et al. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*. 1989;76(1):117–123.
24. Rubin DB. The Bayesian bootstrap. *Ann Stat*. 1981;9(1):130–134.
25. Borgan O, Langholz B, Samuelsen SO, et al. Exposure stratified case-cohort designs. *Lifetime Data Anal*. 2000;6(1):39–58.
26. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scand J Stat*. 2007;34(1):103–119.
27. Hornung RW, Meinhardt TJ. Quantitative risk assessment of lung cancer in US uranium miners. *Health Phys*. 1987;52(4):417–430.
28. Langholz B, Thomas D, Xiang A, et al. Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *Am J Ind Med*. 1999;35(3):246–256.
29. Stram DO, Langholz B, Huberman M, et al. Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau Uranium Miners cohort. *Health Phys*. 1999;77(3):265–275.
30. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Stat Sci*. 1996;11(3):189–212.