Estimation of absolute risk from nested case-control data

Bryan Langholz
Department of Preventive Medicine,
University of Southern California, School of Medicine,
1540 Alcazar Street, CHP-220, Los Angeles, California 90033, U.S.A.

Ørnulf Borgan
Institute of Mathematics, P.O. Box 1053 Blindern,
University of Oslo, N-0316 Oslo, Norway

9 February 1996

Abstract

Benichou and Gail (1995, *Biometrics* **51**, 182-194) describe methods for estimating the absolute risk of developing disease given a set of covariate values over a given time interval from a case-control study within a cohort. The methods are most suitable for unmatched case-control studies and are restricted to time-fixed categorical covariates. Expanding on methods for estimating relative mortality from nested case-control studies given in Borgan and Langholz (1993, *Biometrics* **49**, 593-602), we show how to estimate absolute risk from individually matched nested case-control data. These methods accommodate continuous and time-dependent covariate histories, the sampling of cases, and various control sampling designs.

1 Introduction

Estimation of absolute risk from case-control studies is an important, but surprisingly neglected area of biostatistical research. Drs. Benichou and Gail have long recognized this and seriously addressed this issue. Their most recent paper, Benichou and Gail (1995), which we will refer to as BG, describe methods for estimating absolute risk from case-control studies under a variety of sampling plans. In their introduction, our work on the estimation of relative mortality from nested case-control studies (Borgan and Langholz, 1993), which we will refer to as BL, is cited and they state that our methods "are not appropriate for the study of the absolute risk of breast cancer, because external populations would be composed of women with various levels of exposure to breast cancer risk factors." In fact, the methodology presented in that paper is quite well suited to address the problem of absolute risk estimation from individually matched nested case-control studies such as the Breast Cancer Detection Demonstration Project (BCDDP) nested case-control study discussed in BG. However, since the focus of our 1993 paper was on relative mortality, the

[0] *Key words:* Absolute risk; Case-control studies; Cohort studies; Competing risks; Follow-up data; Relative risk; Risk set sampling; Survival analysis; Time-dependent covariates.
[0] *Abbreviated title:* Absolute risk in nested case-control studies.

applicability of the methods to the absolute risk estimation problem would benefit from further elaboration and elucidation. In particular, we will discuss the estimation of the absolute risk of a disease, c_1, in the absence or presence of competing risks, c_2, for a given, possibly continuous, time-dependent exposure history. The estimator of the variance we provide estimates the "complete variance" in that it takes into account the variability due to both the estimation of relative risk parameters $\boldsymbol{\beta}_0$ as well as the estimation of the baseline incidence h_1 and competing risk incidence h_2. We further show the estimators are adapted to handle other control sampling designs, in particular, restricting controls to match the case on a set of factors and sampling of cases. The former requires no change to the simple random sampling estimator, the later simply requires a change of "weights" in the estimator. The methods are based on the estimators of cumulative hazard and survival probabilities from the proportional hazards model, the theory for full cohort estimators are reviewed in Andersen, Borgan, Gill and Keiding (1993). We illustrate the methods by estimating the risk of lung cancer given potential radon and smoking histories from a nested case-control study of Colorado Plateau uranium miners.

2 Notation

Following the notation of BG, let $h_1(t;\mathbf{x})$ and $h_2(t;\mathbf{x})$ be, respectively, the cause-specific hazards of c_1 and c_2 for an individual with age t and (possibly) time-dependent covariates $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))$. The hazard h_1 is related to $\mathbf{x}(t)$ through the relationship

$$h_1(t;\mathbf{x}) = h_1(t) r(\boldsymbol{\beta}_0, \mathbf{x}(t)), \tag{1}$$

where $h_1(t)$ is the baseline cause-specific hazard for c_1 and $r(\boldsymbol{\beta}_0, \mathbf{x}(t))$ is the relative risk for an individual with covariates $\mathbf{x}(t)$. Here $\boldsymbol{\beta}_0$ is a vector of unknown parameters and r a relative risk function with $r(\boldsymbol{\beta}; \mathbf{0}) = 1$ for $\boldsymbol{\beta}$ in the parameter space. For the special choice $r(\boldsymbol{\beta}_0, \mathbf{x}(t)) = \exp(\mathbf{x}(t)^{\mathsf{T}}\boldsymbol{\beta}_0)$, formula (1) gives the usual Cox regression model (Cox, 1972). Further we assume that h_2 is independent of $\mathbf{x}(t)$, i.e. $h_2(t;\mathbf{x}) = h_2(t)$.

In BL, as a "special case" of the relative mortality model that was the focus of that paper, we provided methods for estimating the cause-specific integrated hazard of c_1 between times s and t corresponding to an individual with a specified covariate history $\mathbf{x}_0(u)$; $s < u \leq t$; i.e.

$$H_1(s,t;\mathbf{x}_0) = \int_s^t r(\boldsymbol{\beta}_0; \mathbf{x}_0(u)) h_1(u) \mathrm{d}u, \tag{2}$$

in the special case where $\mathbf{x}_0$ is fixed over time. This is the expected number of events during $(s,t]$ per subject with covariate history $\mathbf{x}_0(u)$ assuming that subjects are observed over the entire time interval so that the event may occur more than once during the interval. Since we are interested in death which only occurs once or the first occurence of disease, $H_1(s,t;\mathbf{x}_0)$ is larger than the risk of disease or death. However, for rare diseases, it is approximately the risk of disease in the absence of competing risk. The "exact" risk equals the probability that at least one event occurs during $(s,t]$. This is given as $1 - S_0(s,t;\mathbf{x}_0)$, where $S_0(s,t;\mathbf{x}_0) = \exp\{-H_1(s,t;\mathbf{x}_0)\}$ is the probability that no event occurs (Day, 1976; Breslow and Day, 1987, Section 2.2).

Absolute risk in the presence of competing risks is the focus of BG. Specifically, we will explore the estimation of the absolute risk of c_1 in the age interval $(s,t]$ in the presence of competing risks c_2 for a person of age s with covariate history $\mathbf{x}_0(u)$; $s < u \leq t$, i.e.

$$\pi(s,t;\mathbf{x}_0) = \int_s^t S(s,u;\mathbf{x}_0) h_1(u;\mathbf{x}_0) \mathrm{d}u \tag{3}$$

with

$$S(s,u;\mathbf{x}_0) = \exp\left\{ -\int_s^u [h_1(v;\mathbf{x}_0) + h_2(v)]\,\mathrm{d}v \right\} \tag{4}$$

being the probability that neither c_1 nor c_2 occur between age s and u. This is the same quantity as (1.1) in BG expressed in a way that is more convenient for our purposes. The exact absolute risk in the absence of competing risks is a special case with $h_2(v) \equiv 0$. ¡

3 Estimation of absolute risk

We consider a well defined cohort, and assume that there are no tied event times. Further we denote by $t_1 < t_2 < \cdots$ the times when an occurrence of either c_1 or c_2 is observed, and let $d_{1j} = 1$ if an occurrence of c_1 was observed at t_j, $d_{1j} = 0$ otherwise. We define d_{2j} similarly.

We do not have complete covariate information. Rather we assume for c_1 that only nested case-control data are available (Thomas, 1977). Thus at each t_j with $d_{1j} = 1$ a random sample of controls of size $m-1$ is selected without replacement from those at risk. We let $\widetilde{\mathcal{R}}(t_j)$ denote the "sampled risk set" at t_j consisting of the case together with its sampled set of controls. Covariate information is then collected on the individuals in the sampled risk sets, but are not needed for the other individuals in the cohort. The number of individuals at risk at any event time t_j is assumed known, however, and is denoted $n(t_j)$.

Estimation of the regression parameter $\boldsymbol{\beta}_0$ in (1) is based on the partial likelihood

$$\mathcal{L}(\boldsymbol{\beta}) = \prod \left\{ \frac{r(\boldsymbol{\beta};\mathbf{x}_{i_j}(t_j))}{\sum_{l\in\widetilde{\mathcal{R}}(t_j)} r(\boldsymbol{\beta};\mathbf{x}_l(t_j))} \right\}. \tag{5}$$

Here the product extends over all t_j with $d_{1j} = 1$, i_j is the case at t_j and $\mathbf{x}_l(t_j)$ the covariate vector for individual l. Along the lines of Borgan, Goldstein and Langholz (1995) it may be proved that $\widehat{\boldsymbol{\beta}}$ is asymptotically multivariate normally distributed around the true value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$. Its covariance matrix may be estimated by $\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}$, where $\mathcal{I}(\boldsymbol{\beta}) = -(\partial^2/\partial\boldsymbol{\beta}^2)\log\mathcal{L}(\boldsymbol{\beta})$ is the observed information matrix.

3.1 No competing risks

BL considered nested case-control sampling for the Cox proportional hazards model and proposed an estimator for the integrated hazard for an individual in the cohort with a

specified covariate $\mathbf{x}_0$, fixed over time. Their argument readily extends to the set-up of the present note. The result is that (2) is estimated by

$$\widehat{H}_1(s,t;\mathbf{x}_0) = \sum_{s<t_j\le t} \widehat{h}_1(t_j;\mathbf{x}_0), \tag{6}$$

where

$$\widehat{h}_1(t_j;\mathbf{x}_0) = \frac{r(\widehat{\boldsymbol{\beta}};\mathbf{x}_0(t_j))}{\sum_{l\in\widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}};\mathbf{x}_l(t_j))w_l(t_j)} d_{1j}. \tag{7}$$

with $w_l(t_j) = n(t_j)/m$. In order to estimate the variance of (6), let $\widehat{\omega}^2(s,t;\mathbf{x}_0) = \sum_{s<t_j\le t} \widehat{h}_1(t_j;\mathbf{x}_0)^2$ and $\widehat{\mathbf{B}}(s,t;\mathbf{x}_0) = \sum_{s<t_j\le t} \widehat{\mathbf{b}}(t_j;\mathbf{x}_0)$, where

$$\widehat{\mathbf{b}}(t_j;\mathbf{x}_0) = \tag{8}$$

$$\left\{ \frac{\dot{\mathbf{r}}(\widehat{\boldsymbol{\beta}};\mathbf{x}_0(t_j))}{\sum_{l\in\widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}};\mathbf{x}_l(t_j))w_l(t_j)} - \frac{r(\widehat{\boldsymbol{\beta}};\mathbf{x}_0(t_j)) \sum_{l\in\widetilde{\mathcal{R}}(t_j)} \dot{\mathbf{r}}(\widehat{\boldsymbol{\beta}};\mathbf{x}_l(t_j))w_l(t_j)}{\left\{\sum_{l\in\widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}};\mathbf{x}_l(t_j))w_l(t_j)\right\}^2} \right\} d_{1j}$$

with $\dot{\mathbf{r}}(\boldsymbol{\beta};\mathbf{x}(u)) = \frac{\partial}{\partial\boldsymbol{\beta}} r(\boldsymbol{\beta};\mathbf{x}(u))$. Arguing along the lines of Borgan, Goldstein and Langholz (1995) one may show that $\widehat{H}_1(s,t;\mathbf{x}_0)$ is asymptotically normally distributed around $H_1(s,t;\mathbf{x}_0)$, with variance that may be estimated by

$$\widehat{\mathrm{Var}}(\widehat{H}_1(s,t;\mathbf{x}_0)) = \widehat{\omega}^2(s,t;\mathbf{x}_0) + \widehat{\mathbf{B}}(s,t;\mathbf{x}_0)^{\mathsf{T}} \mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1} \widehat{\mathbf{B}}(s,t;\mathbf{x}_0).$$

The $\widehat{\omega}^2$ term is due to the variability in estimating the hazard while the second term accounts for the variability due to the estimation of the relative risk parameters $\boldsymbol{\beta}_0$. These estimators correspond to equations (14) and (15) in BL for the average relative mortality over a time interval and its variance.

The exact risk, in the absence of competing risks, is estimated as one minus the Kaplan-Meier-type estimator

$$\widehat{S}_0(s,t;\mathbf{x}_0) = \prod_{s<t_j\le t} \left(1 - \widehat{h}_1(t_j;\mathbf{x}_0)\right).$$

Further its variance may be estimated as $\widehat{S}_0^2(s,t;\mathbf{x}_0)\widehat{\mathrm{Var}}(\widehat{H}_1(s,t;\mathbf{x}_0))$ (e.g. Andersen, Borgan, Gill and Keiding, 1993; pp. 509-510). These estimators are valid even for "common diseases."

3.2 With competing risks

The absolute risk (3) of c_1 in the presence of competing risks c_2 may be estimated by

$$\widehat{\pi}(s,t;\mathbf{x}_0) = \sum_{s<t_j\le t} \widehat{S}(s,t_{j-1};\mathbf{x}_0)\widehat{h}_1(t_j;\mathbf{x}_0), \tag{9}$$

where

$$\widehat{S}(s,u;\mathbf{x}_0) = \prod_{s<t_j\le u} \left(1 - \widehat{h}_1(t_j;\mathbf{x}_0) - d_{2j}/n(t_j)\right)$$

is the Kaplan-Meir-type estimator of (4).

The absolute risk estimator (9) is of the Aalen-Johansen-type; cf. Andersen, Borgan, Gill and Keiding (1993; Section IV.4, in particular pp. 298–299). Therefore, combining the arguments of Andersen, Borgan, Gill and Keiding (1993; pp. 512–515) on Aalen-Johansen type estimators with those of Borgan, Goldstein and Langholz (1995) on baseline hazard estimation for nested case-control data, one may prove that $\widehat{\pi}(s,t;\mathbf{x}_0)$ is asymptotically normally distributed around (3) with variance that may be estimated by

$$\widehat{\mathrm{Var}}(\widehat{\pi}(s,t;\mathbf{x}_0)) = \widehat{\mathrm{Var}}_1(\widehat{\pi}(s,t;\mathbf{x}_0)) + \widehat{\mathrm{Var}}_2(\widehat{\pi}(s,t;\mathbf{x}_0)).$$

Here

$$\begin{aligned}\widehat{\mathrm{Var}}_2(\widehat{\pi}(s,t;\mathbf{x}_0)) &= \sum_{s<t_j\le t} \left[\widehat{S}(s,t_j;\mathbf{x}_0)\left\{1-\widehat{\pi}(t_j,t;\mathbf{x}_0)\right\}\widehat{h}_1(t_j;\mathbf{x}_0)\right]^2 d_{1j} \\ &\quad + \sum_{s<t_j\le t} \left\{\widehat{S}(s,t_j;\mathbf{x}_0)\widehat{\pi}(t_j,t;\mathbf{x}_0)/n(t_j)\right\}^2 d_{2j},\end{aligned}$$

and

$$\widehat{\mathrm{Var}}_1(\widehat{\pi}(s,t;\mathbf{x}_0)) = \widehat{\mathbf{G}}(s,t;\mathbf{x}_0)^{\mathsf{T}}\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\widehat{\mathbf{G}}(s,t;\mathbf{x}_0),$$

with

$$\widehat{\mathbf{G}}(s,t;\mathbf{x}_0) = \sum_{s<t_j\le t} \widehat{S}(s,t_j;\mathbf{x}_0)\left\{1-\widehat{\pi}(t_j,t;\mathbf{x}_0)\right\}\widehat{\mathbf{b}}(t_j;\mathbf{x}_0)d_{1j}.$$

The estimator of exact absolute risk of c_1 with no competing risks as well as the estimator for its variance (both given in the preceding subsection) are obtained as special cases by setting $d_{2j} = 0$ in all the expressions above.

3.3 Other sampling designs

The estimators described above are easily generalizable to a wide range of nested case-control sampling designs. All that is required is to use the $w_l(t_j)$ appropriate to the sampling design in (7) and (8) and to insert the same weights in the partial likelihood (5). This flexibility allows us to address a number of complications in the BCDDP nested case-control study. First, controls were randomly sampled from those who matched the case on five factors. This is done to control for the confounding effect of those factors when estimating relative risk using the partial likelihood (5), we desire a risk estimate that "pools" over these factors. This situation is directly accommodated by the methods given in Borgan, Goldstein and Langholz (1995). Briefly, we assume a common baseline hazard across the matching factors. The controls, though, are restricted to random sampling of

subjects within matching factor cells. Because the sets of possible controls occur with equal probability, it is easy to see from equation (4.2) in Borgan, Goldstein and Langholz (1995) that the appropriate weights are the same as those for random sampling from all at risk controls, $w_l(t_j) = n(t_j)/m$. Thus, the absolute risk and variance estimators of the last section are all valid without change.

Another feature of the BCDDP nested case-control study is the sampling of cases. We assume that a subject i is sampled as a Bernoulli coin flip with probability $\rho_i(t)$ if i becomes a case at time t. Estimation of the relative risk parameters $\boldsymbol{\beta}_0$ using a weighted partial likelihood is discussed in Langholz and Borgan (1995). The first of two situations we will consider corresponds to that of the BCDDP study. Suppose that covariate information is only available for cases in the case-control study. In this situation, d_{1j} is set to one only for cases included in the case-control study and the appropriate weights are $w_l(t_j) = \rho_l(t_j)n(t_j)/m$. If all cases are sampled with equal probability, the denominator of (7) is just the average of the relative risks in the case-control set "boosted up" by a factor that is the numbers at risk that would be expected to yield the number of cases used in the case-control study. The second situation is where covariate information is available on all cases but not all cases have controls. In this situation d_{1j} is one for all cases and the appropriate weights are $w_l(t_j) = \rho_l(t_j)n(t_j)/\sum_{k\in\widetilde{\mathcal{R}}(t_j)}\rho_k(t_j)$ for members of case-control sets and $n(t_j)$ for single cases. Since the sampling probabilities may be subject dependent, these weightings accommodate stratified sampling of cases by setting ρ_i equal to the probability associated with i's stratum, a situation discussed in BG.

4 Example

We illustrate these methods using a 1:1 nested case-control sample from a cohort of uranium miners from the Colorado Plateau. This cohort was assembled to study the effects of radon exposure and smoking on mortality rates and has been described in detail in earlier publications (e.g. Lundin, Wagoner, and Archer, 1971; Hornung and Meinhardt, 1987). Lung cancer mortality was taken to be c_1 and other causes of death as c_2. The cohort consists of 3,347 Caucasian male miners recruited between 1950 and 1960 and was traced for mortality outcomes through December 31, 1982, by which time there were 258 lung cancer deaths and 2087 deaths from other causes. Exposure data included radon exposure, in working level months (WLM) (Committee on the Biological Effects of Ionizing Radiation, 1988, p. 27), and smoking histories, in number of packs of cigarettes (20 cigarettes per pack) smoked per day.

We consider age as the basic time scale and, as there has been a well known secular trend in lung cancer rates in the general United States population, calendar time was treated as a matching factor with levels defined as the six five year periods 1950-1954, 1955-1959,...,1975-1979, and 1980-1982. Although covariate information is available on all cohort subjects, in order to illustrate the methods each case (tied failure times were randomly broken) was matched to a single control randomly sampled from subjects who were on study at the case's age of death and in the same calendar period as the case at that age. (We note that even when full cohort data is available it may be useful to use a nested case-control sample to reduce error checking and the computational burden (Whittemore and McMillan, 1983; Thomas *et al.*, 1994).)

Table 1: Risk (95% confidence interval), in percent, of lung cancer death with specific radon and smoking histories during ages 40-49, 50-59, and 60-69. Based on the fitted values for 1:1 case-control data set with a cumulative radon and smoking model[a].

Radon exposure					
Age start	Duration (years)	Total dose (WLM)	Smoking[b] (packs/day)	No competing risks[c]	With competing risks
			Age interval 40-49		
–	–	0	0	0.16 (0.06-0.42)	0.14 (0.05-0.37)
–	–	0	0.5	0.3 (0.2-0.7)	0.3 (0.1-0.6)
20	30	480	0.5	0.8 (0.6-1.2)	0.7 (0.5-1.1)
20	30	960	1.0	2.0 (1.4-2.9)	1.7 (1.2-2.5)
			Age interval 50-59		
–	–	0	0	0.4 (0.2-1.0)	0.3 (0.1-0.8)
–	–	0	0.5	1.0 (0.5-2.0)	0.8 (0.4-1.5)
20	30	480	0.5	2.9 (2.3-3.8)	2.3 (1.8-2.9)
20	30	960	1.0	7.7 (5.7-10.5)	5.9 (4.4-7.9)
			Age interval 60-69		
–	–	0	0	0.6 (0.2-1.6)	0.5 (0.2-1.2)
–	–	0	0.5	1.7 (0.9-3.5)	1.3 (0.6-2.5)
20	30	480	0.5	5.2 (3.9-7.0)	3.8 (2.9-5.0)
20	30	960	1.0	14.3 (10.1-20.2)	10.0 (7.4-13.4)

[a] $r(\boldsymbol{\beta}; \mathbf{x}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t))$.
[b] Smoking assumed to start at age 20 and continue throughout life.
[c] Cumulative hazard estimator.

We summarized radon and smoking data into cumulative exposure up to two years prior to the age of death of the case. Thus, we consider as covariates $\mathbf{x}(t) = (R(t), S(t))$, where $R(t)$ is cumulative radon exposure measured in working level months (WLM) up to two years prior to age t, and $S(t)$ is cumulative smoking in number of packs smoked up to two years prior to age t. The relative risk was modeled with radon and smoking acting multiplicatively with

$$r(\beta, \mathbf{x}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t)).$$

The parameter estimates from the 1:1 case-control study were $\widehat{\beta}_R = 0.42$ per 100 WLM (s.e. $= 0.20$), and $\widehat{\beta}_S = 0.23$ per 1000 packs of cigarettes (s.e. $= 0.10$), with a correlation between the estimates of 0.15.

We then computed the predicted absolute risk for a given radon exposure history with constant exposure intensity described by the age start of exposure, duration of exposure, and total exposure. Thus, radon exposure $R(t)$ is zero up to $t - 2 =$ the age at start

of exposure, then increases linearly up to the total exposure at $t - 2 =$ the age start of exposure plus duration of exposure, and is constant at the total exposure thereafter. Smoking was described by the number of packs per day and we assumed that smoking began at age 20 and continued throughout life at the same level. The predicted risk of lung cancer for various exposure histories during ages 40-49, 50-59, and 60-69 are given in Table 1 with confidence intervals computed using the log-transform (Bie, Borgan and Liestøl, 1987) in the same way as in BL for relative mortality. (Confidence intervals based on other transformations may be more appropriate when risks are "large;" e.g. Kalbfleisch and Prentice, 1980; Borgan and Liestøl, 1990.) The no competing risks estimates are based on the cumulative hazard estimator (6) and are interpreted as estimating the probability of death due to lung cancer in those for whom the only reason they would have died during that age interval would be because of lung cancer. As discussed in Section 2, the cumulative hazard estimator is an overestimate of risk. However, comparing the estimates in the fifth column to those based on the exact risk estimator described at the end of Section 3.1 (not shown), the differences in the estimates for predicted risks of less than 5% were small enough to be equal at the two significance digits reported. There were slight differences in predicted risks greater than 5%. Not surprisingly, the largest difference for values in the table were for the last row, where the exact risk estimate was 13.5% (CI=9.8-18.5) compared to cumulative hazard estimate of 14.3% (CI=10.1-20.2). The last column gives the predicted risks of lung cancer in the presence of competing risks, the risk of lung cancer during the age interval given the pattern of deaths from other causes for the uranium miners in this study. These estimates make use of the ages at death and numbers at risk, but no radon or smoking information, from the 2087 miners who died of causes other than lung cancer. These values are, of course, less than the no competing risks risk estimates, the differences between them increases with age as the chance of dying from other causes increases.

5 Discussion

The methods for cumulative hazard estimation from nested case-control data proposed in BL are the natural extention of semi-parametric methods developed for the full cohort. In this report, we have presented absolute risk estimators from nested case-control data, with or without competing risks, again parallel those appropriate for the full cohort, and have generalized the methods to accommodate time-dependent exposure histories. Given the generality of the applications, the estimators and associated variances are relatively simple to compute. All the components are based on simple sums or products of information from the case-control study, the estimated relative risk parameters, and the numbers at risk at the failure times. Further, these estimators are easily adapted to accommodate sampling of cases and control selection other than simple random sampling. The appropriate weights for counter-matching (Langholz and Borgan, 1995) and other sampling designs for the controls are given in Borgan, Goldstein and Langholz (1995).

The estimators presented here were developed specifically for nested case-control data where cases are individually matched to controls who were alive and on study at the failure's time of death or disease. This is the situation in the BCDDP nested case-control study and it would appear that our methods would be more appropriate for estimation

of absolute risk from that study than those presented in BG. Their methods are clearly better suited for unmatched case-control studies; application to failure time data requires a series of "rare disease" approximations, breaking the matching in order to estimate risk, and restriction to categorical covariates. Further, our methods include standard estimators for the full cohort as a special case (weights are set equal to one for each risk set member) while the BG methods do not yield the estimator for unmatched full cohort data. In addition to the basic case-control data, the BG estimators only use the overall incidence rates for the cohort while ours use the numbers at risk at each failure time $n(t_j)$. Thus, we conjecture that our estimators are more efficient since they use more information. On the other hand, the BG estimators have wider applicability. In particular, if appropriate incidence or mortality rates are available, the absolute risk from population based case-control studies may be estimated. We have not explored whether this strategy may be incorporated into our methods.

Acknowledgements

This work was supported by National Cancer Institute grant CA14089.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Verlag, New York. .

Benichou, J. and Gail, M.H. (1995). Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* **51**, 182–194.

Bie, O., Borgan, Ø., and Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, **14**, 221–233.

Borgan, Ø., Langholz B., and Goldstein L. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics (to appear, October issue).*

Borgan, Ø. and Langholz, B. (1993). Non-parametric estimation of relative mortality from nested case-control studies. *Biometrics* **49**, 593-602.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics*, **17**, 35–41.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research.* Volume II – *The Design and Analysis of Cohort Studies*, *IARC scientific publications*, Vol. 82. International Agency for Research on Cancer, Lyon.

Committee on the Biological Effects of Ionizing Radiation (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters, BEIR IV*, National Academy Press, Washington, D.C.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, 187–220.

Day, N. (1976). A new measure of age standardized incidence, the cumulative rate. In *Cancer Incidence in Five Continents, Vol. III (IARC Scientific Publications No. 15)*, (ed. J. Waterhouse, C. Muir, P. Correa, and J. Powell), pp. 443–452. International Agency fo Research on Cancer, Lyon.

Hornung, R. and Meinhardt, T. (1987). Quantitative risk assessment of lung cancer in U. S. uranium miners. *Health Physics*, **52**, 417–430.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82**, 69-79.

Lundin, F., Wagoner, J., and Archer, V. (1971). Radon daughter exposure and repiratory cancer, quantitative and temporal aspects. Joint Monograph 1, U.S. Public Health Service, Washington, D.C.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society A*, **140**, 469–491.

Thomas, D., Pogoda, J., Langholz, B., and Mack, W. (1994). Temporal modifiers of the radon-smoking interaction. *Health Physics*, **66**, 257–262.

Whittemore, A. and McMillan, A. (1983). Lung cancer mortality among U.S. uranium miners: A reappraisal. *Journal of the National Cancer Institute*, **71**, 489–499.