Statistical Science
1996, Vol. 11, No. 1, 35–53

Risk Set Sampling in Epidemiologic Cohort Studies

Bryan Langholz and Larry Goldstein

Abstract. Recent work has extended the methods for the analysis of nested case-control studies to accommodate a broad variety of risk set sampling designs. These results have implications for the design of sampled epidemiologic cohort studies. We describe a model which is a natural extension of the Cox proportional hazards model and may be used to estimate parameters from sampled risk set data. We illustrate how these techniques may be used to solve three diverse design and analysis problems from epidemiologic research.

Key words and phrases: Survival analysis, cohort sampling, case-control studies, martingale, censoring, efficiency.

1. INTRODUCTION

In 1982, a lawsuit was brought against an aircraft manufacturing firm in San Diego County, California, by the families of a group of employees who had died of esophagus cancer. Their claim was that occupational exposures to carcinogenic agents were to blame for the disease. In order to objectively assess whether there truly was an unusually high rate of cancer among employees and whether any particular substances used at the firm were related to the risk of esophagus cancer, a cohort mortality study and a case-control study were undertaken. Investigators compiled basic information such as birth dates, sex, race and dates of employment for the cohort of 14,067 employees from company records. Other data sources were required to determine vital status and, if dead, the cause of death.

For the mortality study, the number of deaths in the cohort from various causes were then compared to those expected based on United States cause-specific mortality rates. There were 14 cases of esophagus cancer observed compared to 12.27 expected (standardized mortality ratio (SMR) = O/E = 1.14, 95% confidence interval 0.62–1.92; Garabrant, Held, Langholz and Bernstein, 1988). Although this excess is certainly not significantly different from 1, cancer mortality was generally lower in the cohort with an SMR for all cancers of 0.84; perhaps a manifestation of the "healthy worker effect." Thus, while this analysis showed that there was no evidence of a cancer outbreak, it did not rule out the possibility of a real excess of esophagus cancer at the plant. A much more informative approach would be to see if particular occupational exposures were associated with disease risk. Job titles and plant location histories for each employee as well as records on substances used in the manufacturing processes were available from company archives, so, in principle, probable employee exposure histories could be generated. However, this would be an exceedingly expensive task to do with any accuracy for all 14,067 subjects. This is why the nested case-control study was undertaken. For each of the 14 cases, four controls matched on year of birth, sex and race, were randomly selected from those who were alive at the case's age of death. Exposure histories, as described above, were further refined through interviews with supervisors and co-workers of the sampled subjects. After an extensive analysis of exposures, an association was found with a single building at the plant, but no specific occupational exposures could be implicated. (In fact, in the "high risk building," office workers were over-represented among the cases.) An out-of-court settlement was eventually reached.

The case-control substudy of the cohort of aircraft manufacturing employees is an example of sampling from the risk set. The risk set associated with a given esophagus cancer case includes the case and "controls," all subjects of the same sex, race and year of birth as the case, and who were alive at and had

Bryan Langholz is an Associate Professor in the Department of Preventive Medicine, University of Southern California, Los Angeles. Larry Goldstein is an Associate Professor in the Department of Mathematics, University of Southern California.

been employed at the firm by the age of death of the case. Instead of using all possible controls in the risk set, four were randomly sampled without replacement.

The nested case-control design has been used in many studies to avoid collection of exposure and other information for the full cohort (e.g., Adelhardt, Møller Jensen and Sand Hansen, 1985; Boice, Blethner, Kleinerman, et al., 1987; Garabrant, Held, Langholz, Peters and Mack, 1992; Kogevinas et al., 1995) or to reduce the computational burden in data analysis (e.g., Liddell, McDonald and Thomas, 1977; Thomas, Pogoda, Langholz and Mack, 1994). In fact, as has long been recognized, most matched case-control studies, ubiquitous in epidemiologic research, are nested case-control studies where the cohort is a (perhaps not well defined) population in a geographic area (Prentice and Breslow, 1978). With $m - 1$ the number of sampled controls, the efficiency of this design relative to the full cohort for testing for an association between a single factor and disease is $(m - 1)/m$ (Breslow and Patton, 1979). Thus, the sample of 70 subjects, or 0.5% of the total cohort, used in this nested case-control study of the aircraft manufacturing firm, provided about 4/5 = 80% efficiency relative to the full cohort for testing associations between *single* exposures and disease.

However, it is becoming rare that the goals of epidemiologic studies can be addressed by simple tests of associations and many more controls may be required to achieve efficiencies this high when the analyses are more complex (Breslow, Lubin, Marek and Langholz, 1983). As the questions considered grow in complexity and the costs involved increase, it becomes advantageous to adapt the sampling designs to take into account the goals of the study and the costs associated with data collection. This has been demonstrated in evaluations of new sampling designs for grouped binary outcome data (e.g., Breslow and Cain, 1988; Weinberg and Wacholder, 1993). Preceded by theoretical developments that linked the analysis of risk set sampled data to those based on the Cox proportional hazards model for the full cohort (Borgan, Goldstein and Langholz, 1995), it is only recently that analogous designs have been developed for the sampling of risk sets. In Section 5, we explore new risk set sampling designs to solve two epidemiologic design problems where exposure information is available on cohort members and a sample is to be drawn to obtain further information. These designs use the exposure information in the sampling of controls and are more efficient than random sampling for the problems they address. One problem we consider is motivated by the need to collect smoking information on a sample of a uranium miners cohort to supplement the investigation of radon exposure, available on all cohort members, and lung cancer rates. In the other design problem, we explore two-stage design options for subsampling a (nested) case-control study of electromagnetic fields and leukemia to investigate a new measure of the fields that is believed to be more relevant to leukemia risk.

The new methods also solve some outstanding analysis problems. In Problem 2 of Section 5, we estimate the risk of lung cancer given continuous and time-dependent radon and smoking histories from nested case-control data. Before tackling these problems, in Section 2, we describe the basic elements of cohort data and introduce the terminology related to nested case-control studies. Section 3 gives a short history of the statistical methods for nested case-control studies and the connection to the partial likelihood approach to the analysis of cohort data. After showing how Midzuno's procedure—a "survey sampling" method for unbiased ratio estimation—provides some intuition for the standard matched data approach to the analysis of nested case-control data, we give the counting process formulation of the problem and describe the estimators of relative and absolute risk parameters that are natural extensions of the full cohort estimators.

2. TERMINOLOGY

The study of Garabrant, Held, Langholz and Bernstein (1988) contains many of the essential ingredients common in epidemiologic work, where questions arise involving the comparison of the incidence of rare diseases given specific risk factors. Investigations of incidence involve a study population, called a *cohort*, followed over some period of time, where disease incidence is compared across subgroups defined by risk factors and other *covariates*. Figure 1 diagrams the basic features of a small hypothetical cohort of 15 subjects. Each subject enters the study at some *entry time*, is *at risk*, denoted by the horizontal line, over some time period, and exits the study at some *exit time*. We assume that there are two reasons for exit. A subject may contract or die from the disease of interest, and thus be a *failure* (represented by "•" in Figure 1) or be *censored*, that is, be alive at the end of the study, died never having had the disease of interest or be lost to followup.

The choice of time scale will depend on the goals of the investigation. Often there is a natural time scale such as time since treatment in intervention or clinical trials, but in many situations the choice

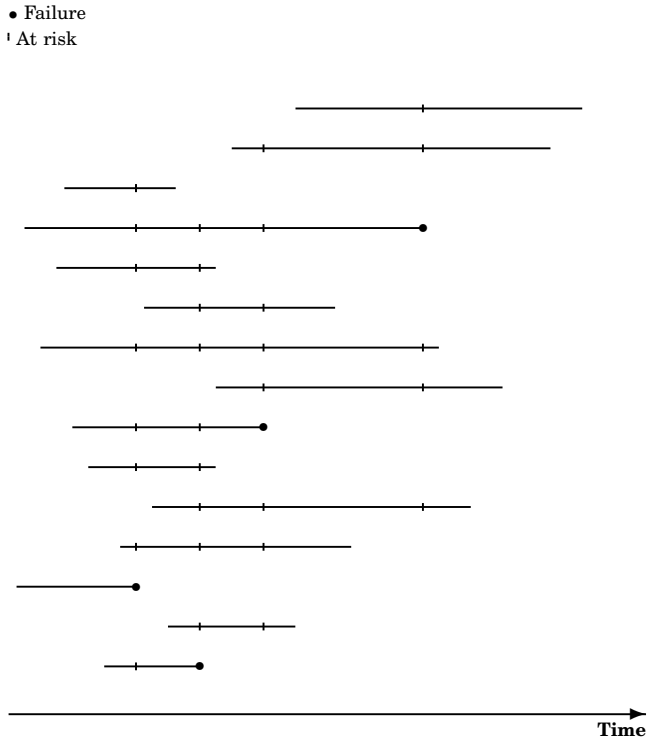

FIG. 1. *Hypothetical cohort. Each line represents a subject's time on study.*

may be less straightforward. In occupation cohort studies, age is often chosen because disease incidence rates vary greatly with age, but calendar time or time since first employment might also be considered (Breslow and Day, 1987; Clayton and Hills, 1993).

Associated with each subject is a *covariate history*, which usually includes factors which are known or believed to be related to the rate of the disease of interest. Covariates may be fixed over time, such as country of origin, race, or gender. Alternatively, they may be time dependent, such as cumulative packs of cigarettes smoked or time since last mammograph. In principle, the covariate value for a subject at a given time should reflect the covariate history up to that time. Covariate information may be classified in many ways, but, for this discussion, it is useful to divide them generically into *exposures*, the covariates we care about, and *confounders*, covariates that we need to consider because they are related to both exposure and disease rates, but would have been held constant in a controlled experiment (Clayton and Hills, 1993, page 135). We also consider *interactions* or *effect modifiers* that quantify how the effect of one covariate may depend on the value of another.

Risk set sampling designs are intrinsically related to semiparametric estimation methods for parameters in the Cox proportional hazards model used in the analysis of full cohort data. In this method, at each failure time a *risk set* is formed that includes the *case*, the failure at that failure time and all *controls*, any other cohort members who are at risk at the failure time (these are denoted by a "|" in Figure 1).

A *sampled risk set* of size m is a subset of the risk set that contains the case and $m - 1$ sampled controls. So, for instance, for 1:1 simple nested case-control sampling, each sampled risk set consists of the case and one control randomly sampled from all the controls in the risk set.

In order to properly assess how disease rates change with level of exposure, control for the effect of relevant confounders is necessary. This may be achieved by modeling the effect of the confounder or by restricting the risk set to those who have the same (or similar) confounder values. We will call the latter procedure *matching* and call confounders treated in this way in the analysis *matching factors*. Further, we will assume that matching factors are categorical. This corresponds to *stratification* of the Cox model. When there are matching factors, the sampled risk set will be a subset of the risk set defined by the failure time and any matching factors. For instance, in the Garabrant, Held, Langholz and Bernstein (1988) study, the matching factors were year of birth, gender and race, and the risk set, from which the four controls were selected, consisted of subjects who matched the case on these factors.

3. STATISTICAL METHODS

The approach which organizes cohort data by risk sets (see Figure 1) leads to a data set which looks just like a matched case-control study (Miettinen, 1969; Breslow and Day, 1987). As described in Cox's landmark 1972 paper (Cox, 1972), this motivates the use of the conditional logistic likelihood for the analysis of such data. The contribution from a case-control set is the conditional probability that the observed case is diseased given that one of the subjects in the case-control set is diseased. With p_k the unconditional probability that subject k becomes diseased, this probability is given by

$$(1) \qquad p_{\text{case}} \Big/ \sum_{\text{case-control set}} p_k.$$

The conditional logistic likelihood is now formed by taking products of the factors (1) over all matched sets. Applying this idea to cohort data, let the incidence of disease depend on a vector of covariate

history summaries $z(t)$ and have a proportional hazards form

$$\lambda(t; z(t)) = \lambda_0(t) r(z(t); \beta_0), \tag{2}$$

where $r(z; \beta_0)$ is the relative risk of disease for an individual with covariates z and $r(0; \beta) = 1$, so $\lambda_0(t)$ is the rate of disease in subjects with $z = 0$. Now consider a single risk set as in Figure 1. The "instantaneous probability" of failure for subject k in the risk set formed at failure time T is $p_k = \lambda(T, Z_k(T))\, dt$, where $Z_k(t)$ is the covariate of individual k. Cancelling the common factor $\lambda_0(T)$ yields a contribution to the conditional logistic likelihood of the form

$$r_{\text{case}} \Big/ \sum_{k \in \mathscr{R}} r_k, \tag{3}$$

where $\mathscr{R}$ is the "case-control set" (all risk set members) and r_k is the relative risk associated with the covariate history for subject k at time T.

The construction of a conditional logistic "likelihood" by analogy to matched case-control studies is heuristic since it is formed as though the matched sets are independent, and this is clearly not the situation for the risk sets. As illustrated in Figure 1, the same subjects may appear in multiple (many!) risk sets. This quantity was eventually termed a "partial likelihood" because it essentially leaves out the information between the failure times (Cox, 1975; Holford, 1976).

With the concept of risk sets as case-control sets, the idea of sampling controls seemed quite natural, again by analogy to case-control study methodology. The efficiency of $m-1$ controls relative to an infinite number of controls when there is a single covariate which is not related to the risk of disease ($\beta_0 = 0$) is $(m-1)/m$ (Ury, 1975). Thus, it was natural to conclude that a random sample of a relatively small number of controls from the risk set would be an efficient way to obtain a sample from the cohort. This approach was first implemented for a cohort of asbestos miners in Quebec (Liddell, McDonald and Thomas, 1977). Four controls were randomly sampled from each risk set. With this reduced set, it was feasible for the investigators to extensively "clean" the data as well as vastly reduce computing time, relative to the full risk set Cox regression.

In the Appendix to Liddell, McDonald and Thomas (1977), Thomas points out that there are two heuristic approaches to the analysis of the sampled data. The first is to use an unbiased estimator of the denominator of the partial likelihood for the full cohort; that is, weight the control r_k's by $(n-1)/(m-1)$ and r_{case} by 1. Although this approach seems natural, it leads to biased estimation of the parameters. The second approach is based on simply viewing the sample as a case-control study and using the conditional logistic likelihood as described above. With this method, the case is weighted the same as the controls, so that the denominator is

$$r_{\text{case}} + \sum_{\text{controls}} r_k = \sum_{k \in \widetilde{\mathscr{R}}} r_k, \tag{4}$$

where $\widetilde{\mathscr{R}}$ is the sampled risk set (i.e., the case and sampled controls). Thomas (correctly) employed the latter in his analyses. Essentially, the problem with the former is that the "inverse of the mean" does not equal the "mean of the inverse." The likelihood based on the latter "unweighted" denominator (4) was shown to have a partial likelihood interpretation in the same spirit as in the full cohort situation (Oakes, 1981). Oakes' proof is essentially formalized in the counting process approach, which we will describe later. First, it is instructive to demonstrate the validity of the unweighted likelihood using a method from sampling theory.

3.1 Midzuno's Procedure

In standard likelihood theory, the expectation of the score—the first derivative of the log likelihood—has expectation zero at the true parameter. That the unweighted likelihood using denominator (4) has this property is a consequence of Midzuno's procedure (Midzuno, 1952), a technique used to obtain an unbiased estimator of the ratio of means. The score contribution associated with (4) is given by

$$\frac{r'_{\text{case}}}{r_{\text{case}}} - \frac{\sum_{\widetilde{\mathscr{R}}} r'_k r_k}{\sum_{\widetilde{\mathscr{R}}} r_k}, \tag{5}$$

where r' is the derivative of r with respect to β, and $\widetilde{R}$ is the sampled risk set. Since the expectation of the full cohort score is zero, it is sufficient to show that the expectation over sampling of the second term is the full cohort value $\sum_{\mathscr{R}} r'_k r_k / \sum_{\mathscr{R}} r_k$.

Let (x_i, y_i), $i \in \mathscr{R} = \{1, \ldots, n\}$, be a set of values such that $x_i \geq 0$ with $\sum_{i \in \mathscr{R}} x_i > 0$. Let $\overline{x}_{\mathscr{R}} = n^{-1} \sum_{i \in \mathscr{R}} x_i$ with $\overline{y}_{\mathscr{R}}$ similarly defined and suppose that we wish to estimate $\theta_{\mathscr{R}} = \overline{y}_{\mathscr{R}} / \overline{x}_{\mathscr{R}}$ using a sample of m of the n (x, y) pairs. With $\widetilde{\mathscr{R}}$ a random set of indices, first note that if the m indices are the result of simple random sampling, then, generally, $E[\theta_{\widetilde{\mathscr{R}}}] \neq \theta_{\mathscr{R}}$. Midzuno's procedure gives a way to obtain an unbiased estimator:

1. Sample an index according to the "x-size biased distribution," that is, pick an index i according to the probabilities $x_i / \sum_{k \in \mathscr{R}} x_k$.
2. Sample $m-1$ subjects randomly, without replacement from the $n-1$ remaining indices.

With $\widetilde{\mathscr{R}}$ the set of the resulting m sampled indices, one can show that the estimator $\theta_{\widetilde{\mathscr{R}}}$ is unbiased for $\theta_{\mathscr{R}}$.

Applying Midzuno's procedure to the score (5), let $x_i = r_i$ and $y_i = r_i' r_i$. The probability that subject i is the case is given by (3), so Step 1 of the procedure is satisfied by the occurrence of the case. The $m - 1$ randomly sampled controls satisfy Step 2. Thus, the expectation over sampling of the score term is just the full cohort score contribution so that the expectation over cases is zero at the true parameter β_0.

3.2 Counting Process Approach

A formal treatment of risk set sampling is based on specifying an appropriate intensity model. In Andersen and Gill's formulation of the Cox model (Andersen and Gill, 1982), a counting process $N_i(t)$, which "counts" failure occurrences for subject i, and a corresponding intensity process $\lambda_i(t)$ are associated with each subject in the cohort. In order to accommodate sampling from the risk sets, define the counting processes $N_{i,\mathbf{r}}(t)$ which record occurrences of the joint event that i fails *and the set of subjects* $\mathbf{r}$ *serves as the sampled risk set*. The intensity processes corresponding to $N_{i,\mathbf{r}}(t)$ take the form of the product

$$\lambda_{i,\mathbf{r}}(t) = \lambda_i(t)\ \pi_t(\mathbf{r}|i), \tag{6}$$

where $\pi_t(\mathbf{r}|i)$ is the probability of choosing $\mathbf{r}$ as the sampled risk set if subject i were to fail at time t. This approach was described for simple random sampling of controls (which we will henceforth call "simple nested case-control sampling") in Borgan and Langholz (1993). Since $m-1$ controls are randomly sampled, without replacement, from the $n-1$ controls in the risk set at time t, $\pi_t(\mathbf{r}|i) = \binom{n-1}{m-1}^{-1}$ for subsets of the risk set of size m that contain i. In the same spirit as the conditional likelihood approach for the full cohort, the conditional probability that i fails and $\widetilde{\mathscr{R}}$ is picked as the sampled risk set given that one of the $k \in \widetilde{\mathscr{R}}$ fails is

$$\frac{r_i\ \pi_t(\widetilde{\mathscr{R}}|i)}{\sum_{k\in\widetilde{\mathscr{R}}} r_k\ \pi_t(\widetilde{\mathscr{R}}|k)} = \frac{r_i}{\sum_{k\in\widetilde{\mathscr{R}}} r_k}, \tag{7}$$

by cancelling the common factor $\binom{n-1}{m-1}^{-1}$. Hence the "partial likelihood" that results upon taking products of the above factors over all failure times is the same as the case-control likelihood based on (4). As mentioned in Section 3.1, that the expectation of the score at the true parameter β_0 is zero is a consequence of Midzuno's procedure providing an unbiased ratio estimator. For establishing asymptotic properties of the maximum partial likelihood estimator $\widehat{\beta}$, the counting process formulation provides a convenient framework. This connection has been explored for the full cohort in Andersen and Gill (1982) and in the sampling context in Goldstein and Langholz (1992) and Borgan, Goldstein and Langholz (1995).

3.3 Other Risk Set Sampling Designs

The counting process formulation of simple nested case-control sampling provided a probabilistic model for formally establishing the properties of the maximum partial likelihood estimator. However, this approach has much wider implications than the limited goal of putting an established sampling technique on a firm theoretical footing. Other methods for sampling the risk set can be accommodated by specifying appropriate sampling probabilities $\pi_t(\mathbf{r}|i)$. This innovation, explored in some detail in Borgan, Goldstein and Langholz (1995), provides an analysis method for new sampling designs where controls are sampled in a "nonrepresentative" way. In particular, this has enabled our exploration of designs that are solutions to the epidemiologic design problems described in Section 5.

The general formulation in terms of given set sampling probabilities $\pi_t(\mathbf{r}|i)$ often simplifies because of cancellation of common terms; indeed, this cancellation was noted when simplifying (7) for nested case-control sampling. Thus, a likelihood contribution from each sampled risk set is of the form

$$\frac{\pi_t(\mathbf{r}|\text{case})\ r_{\text{case}}}{\sum_{k\in\widetilde{\mathscr{R}}} \pi_t(\mathbf{r}|k)\ r_k} = \frac{(Wr)_{\text{case}}}{\sum_{k\in\widetilde{\mathscr{R}}} (W_k r_k)}, \tag{8}$$

where the W_k are subject (and sampled risk set) specific *risk weights* chosen to be "convenient" multiples of the $\pi_t(\mathbf{r}|k)$ in the partial likelihood analysis. As any multiple of the risk weights lead to the same factor as above, it is often convenient to form the canonical weights

$$W_k' = \frac{\pi_t(\mathbf{r}|k)}{\sum_{j\in\mathbf{r}} \pi_t(\mathbf{r}|j)} = \frac{W_k}{\sum_{j\in\mathbf{r}} W_j}, \tag{9}$$

which are components in the estimators of absolute risk described in the next section. The partial likelihood obtained by taking products of the factors (8) over all failure times has the usual basic properties of a likelihood, and standard conditional logistic regression software used for the analysis of matched case-control studies can be used for data analysis, treating the W as risk weights or "offsets" in the model.

3.4 Risk Estimation

As the previous sections suggest, methods for the estimation of the *relative* risk parameter β_0 from case-control studies have been studied extensively. There has been much less attention to the problem of estimating *absolute* risk, which are functionals of the hazard λ. While relative risk is an important measure for studying disease etiology, absolute risk is important for public health planning and personal risk counciling. Now, in the simple binomial situation, the estimator of the binomial probability is just the number of cases divided by the *total* number in the relevant population. Analogously, in order to estimate absolute risk from failure time data, "denominator" information about the *entire* population at risk (not just a sample) is required. Thus, absolute risk cannot be estimated using only the data in a case-control sample, supplemental information that links the case-control study to the cohort from which it is sampled is required (e.g., Benichou and Wacholder, 1994; Benichou and Gail, 1995). For nested case-control sampling, one approach uses overall cohort disease rates as the link between the sample and the cohort (Benichou and Gail, 1995). Another approach, which we will explore here, uses the number of subjects in each risk set to provide this link (Breslow and Langholz, 1987; Borgan and Langholz, 1993). Appropriate methods for estimation of absolute risk for the general risk set sampling described in the last section are given in Borgan, Goldstein and Langholz (1995) and Langholz and Borgan (1996).

Let z^0 be a covariate history and $r^0 = r(z^0; \beta_0)$ be the relative risk associated with z^0 according to the model. The estimator we have proposed of the cause-specific cumulative hazard (the "risk" for rare diseases) associated with that history between times s and t, that is, an estimator of

$$\int_s^t \lambda_0(u) r(z^0(u); \beta_0)\, du, \tag{10}$$

is a generalization of the Breslow estimator for full cohort data (see discussion in Cox, 1972). With n the number at risk, $\widehat{r}_k = r(z_k, \widehat{\beta})$ the relative risk for individual k predicted using $\widehat{\beta}$ and W'_k as defined in (9), the estimator of (10) is the sum of contributions of the form

$$\widehat{r}^{\,0} \Big/ n \sum_{k \in \widetilde{\mathscr{R}}} (W'_k \widehat{r}_k) \tag{11}$$

over all failure times between s and t. Note that by (9), for the full cohort $W'_k = 1/n$, which cancels the leading n in the denominator of (11). Setting $z^0(t) \equiv 0$, we have the usual Breslow estimator of the baseline cumulative hazard. The estimator just described is "almost unbiased" in the same sense as the Breslow estimator (Andersen, Borgan, Gill and Keiding, 1992, Section VII.2.2). Langholz and Borgan (1996) provide a relatively simple variance estimator which takes the estimation of β_0 into account.

4. COLORADO PLATEAU URANIUM MINERS COHORT

In this section, we first describe the data set we will use to illustrate some of the methods and then give the results from fitting a series of models using the full cohort and simple nested case-control samples. These preliminary analyses will be compared to those based on new risk set sampling designs that we will investigate in Section 5 and are proposed as solutions to specific epidemiologic cohort sampling problems.

4.1 Description of the Data

The Colorado Plateau uranium miners cohort data were collected to study the effects of radon exposure and smoking on the rates of lung cancer and has been described in detail in earlier publications (e.g., Lundin, Wagoner and Archer, 1971; Hornung and Meinhardt, 1987; Lubin et al., 1994). The cohort consists of 3347 Caucasian male miners who worked underground at least one month in the uranium mines of the four-state Colorado Plateau area and were examined at least once by Public Health Service physicians between 1950 and 1960. These miners were traced for mortality outcomes through December 31, 1982, by which time 258 lung cancer deaths had occurred. Entry time into the cohort was the date of first examination or first underground work, whichever came later. Exit time was the date at death, December 31, 1982 if known alive at that time or date last known to be alive if lost to follow-up. Subjects who died of lung cancer were taken to be the failures and all others were censored at their exit times.

Job histories were obtained from each of the mining companies and combined with available data on annual mine radon levels to estimate each miner's occupational annual mine exposures. A smoking history was taken at the first examination and updated with each subsequent exam. The data available included the age of starting and quitting smoking and the number of packs of cigarettes smoked per day. Thus, for any age on study, it is possible to compute summary measures of radon and smoking exposures.

Because smoking and radon information are available for all cohort members, we will be able to

compare the results obtained when sampling the cohort to those obtained using the full cohort.

4.2 Models

In previously published analyses and those used to illustrate the methods here, we consider age as the basic time scale. There has been a well known secular trend in lung cancer rates in the general United States population, so calendar time was treated as a matching factor (Section 2) with levels defined as the six five-year periods 1950–1954, 1955–1959, ..., 1975–1979 and 1980–1982. This matching factor is time dependent in that a subject will change from one matching level to another with age. Thus, the intensity model assumes calendar period-specific baseline hazards, an extension of model (2), where λ_0 is replaced by λ_c, where c is the calendar period.

Radon and smoking data were summarized into simple cumulative dose measures. Since lung cancer victims survive about two years after being diagnosed, and exposures after diagnosis have no effect on the course of the disease, exposures are cumulated only up to two years prior to the case's age of death. [As our emphasis here is on sampling methods, this summary of exposures was chosen for its simplicity, but is not the most realistic or best fitting summary. See Whittemore and McMillan (1983), Thomas, Pogoda, Langholz and Mack (1994) and Lubin et al. (1994) for more realistic exposure models.] Thus, we consider as covariates $\mathbf{Z}(t) = (R(t), S(t))$, 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.

For example, consider a case who dies at age 54.2 in 1964. The risk set for this case consists of all those who are alive and on study at age 54.2 and reach that age during 1960–1964. Furthermore, the covariate $\mathbf{Z}(54.2)$ is computed as cumulative radon and smoking up to age 52.2.

We considered four models of the relative risk as a function of radon and smoking:
Radon:

$$r(\beta, \mathbf{Z}(t)) = 1 + \beta_R R(t); \tag{12}$$

Smoking:

$$r(\beta, \mathbf{Z}(t)) = 1 + \beta_S S(t); \tag{13}$$

Radon, smoking:

$$r(\beta, \mathbf{Z}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t)); \tag{14}$$

Interaction:

$$r(\beta, \mathbf{Z}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t)) \cdot \exp(\beta_{RS} R(t) S(t)). \tag{15}$$

The reasons for choosing these models for analysis are described in detail in Thomas, Pogoda, Langholz and Mack (1994).

4.3 Full Cohort Analysis

The *full cohort* column of Table 1 gives the results of fitting the four models (12)–(15) using the entire risk sets, that is, the full cohort Cox partial likelihood analysis. Fits of models (12) and (13) revealed strong univariate associations between radon and smoking and lung cancer mortality rates. There is not much difference between the univariate estimates of the effect of these exposures after adjustment for each other, indicating that there is little correlation between the two exposures. From model (14), (radon-adjusted) smoking excess relative risk is $\beta_S = 0.17$ per 1000 packs of cigarettes and the (smoking-adjusted) radon excess relative risk is $\beta_R = 0.40$ per 100 WLMs. From the negative estimated interaction parameter β_{RS} from model (15), there is evidence of effect modification between radon and smoking (Wald test $= -0.68/0.27 = -2.5$, two-sided p-value $= 0.01$) with the joint effect of the exposures somewhat less than predicted by the main effects.

4.4 Simple Nested Case-Control Sample Analysis

Simple nested case-control samples with one and three controls were drawn from the risk set established by the case's failure time and five-year calendar period at death. This resulted in samples of 478 and 837 distinct subjects or about 14 and 25% of the cohort for the 1:1 and 1:3 samples, respectively. [Because subjects can be sampled as controls in multiple risk sets and failures can serve as controls in risk sets prior to their failure times (Lubin and Gail, 1984), the number of distinct subjects will be somewhat less than the appropriate multiple of 258, the number of failures.] The results of the analyses of these data sets are given in the third and fourth columns of Table 1. Qualitatively, the parameter estimates are similar to those obtained by the full cohort analysis. Note that the standard errors for the 1:1 random sample are about double those of the full cohort for both the radon and smoking variables. If there were no effect of these exposures, by the "$(m-1)/m$" relative efficiency rule, these standard errors for the 1:1 sample ($m = 2$) would be anticipated to be about $1.4 \approx \sqrt{2}$ that of the full co-

TABLE 1
Parameter estimates (standard errors) for radon and smoking models using all controls, that is, the full cohort, and random and counter-matched sampling of controls from the risk sets

Model[a]	Full cohort	Random sampling		Counter-matching		Hybrid
		1:1	1:3	1:1	1:3	1:1:1
Univariate models						
Radon (β_R)[b]	0.36 (0.10)	0.41 (0.19)	0.41 (0.15)	0.33 (0.11)	0.36 (0.11)	0.35 (0.11)
Smoking (β_S)[c]	0.16 (0.05)	0.18 (0.07)	0.20 (0.06)	0.37 (0.15)	0.23 (0.08)	0.19 (0.07)
Adjusted model[d]						
Radon (β_R)	0.38 (0.11)	0.42 (0.20)	0.43 (0.16)	0.39 (0.14)	0.41 (0.13)	0.44 (0.16)
Smoking (β_S)	0.17 (0.05)	0.23 (0.10)	0.20 (0.07)	0.25 (0.10)	0.19 (0.07)	0.23 (0.09)
Interaction model[e]						
Radon (β_R)	0.67 (0.27)	0.51 (0.29)	0.53 (0.24)	0.54 (0.28)	0.50 (0.21)	0.62 (0.30)
Smoking (β_S)	0.24 (0.08)	0.25 (0.12)	0.22 (0.08)	0.30 (0.13)	0.22 (0.079)	0.29 (0.11)
Interaction (β_{RS})	−0.68 (0.27)	−0.41 (0.70)	−0.41 (0.42)	−0.53 (0.46)	−0.31 (0.36)	−0.53 (0.42)
Number of distinct subjects	3347	478	837	473	765	670

[a] Radon slopes given as per 100 WLM. Smoking slopes given as per 1000 cumulative packs of cigarettes. Interaction slopes given as per 10^8 cumulative WLM∗packs.
[b] Univariate radon: $r(\beta, \mathbf{Z}(t)) = 1 + \beta_R R(t)$.
[c] Univariate smoking: $r(\beta, \mathbf{Z}(t)) = 1 + \beta_S S(t)$.
[d] Adjusted model: $r(\beta, \mathbf{Z}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t))$.
[e] Interaction model: $r(\beta, \mathbf{Z}(t)) = (1 + \beta_R R(t))(1 + \beta_S S(t)) \exp(\beta_{RS} R(t) S(t))$.

hort. With the rather strong effects observed here, the standard errors are somewhat larger, as would be expected based on large sample investigations of this question in some particular situations (Breslow and Patton, 1979; Breslow, Lubin, Marek and Langholz, 1983; Goldstein and Langholz, 1992). The 1:3 sample yields smaller standard errors, but they are still quite a bit larger than those obtained by the full cohort analysis.

5. SOME SPECIFIC PROBLEMS

In this section, we consider a number of study design and analysis problems that we have encountered in our work with epidemiologic data, and show how our sampling methods provide solutions.

PROBLEM 1. *Informative sampling based on exposure information.* Suppose that researchers have assembled a cohort and have collected some exposure-related information for all (or, at least most) of the subjects. It is desired to collect further information on a sample of the cohort, perhaps more precise exposure measurements, confounder information or information on other potential exposures. As an example that we will use throughout this section, suppose that only radon exposure information had been collected for Colorado Plateau uranium miners and smoking histories were to be obtained on a sample to assess the role of smoking as a confounder or effect modifier of radon exposure on lung cancer rates. Random sampling of controls from the risk sets is always a possible design, but it seems wasteful of the exposure information already available on cohort members; the efficiency for estimation of exposure parameters would be the same as if exposure data had been collected only for the sample. Intuitively, a sampling design that increases the variability in exposure values over that of random sampling would be more efficient. This is the principle behind *counter-matching*, an exposure stratified sampling method developed for this situation (Langholz and Borgan, 1995; Langholz and Clayton, 1994). To get a feel for the design and the method of analysis, consider the simple case of a dichotomous exposure. In the 1:1 counter-matching design, a control is randomly sampled from those in the risk set that have exposure status *opposite* of that of the case. (Note that this is the opposite of what is done in matching, as the name indicates.) With n_0 and n_1 the number of unexposed and exposed subjects in the risk set, $\pi(\mathbf{r}|j)$ is n_0^{-1} or n_1^{-1} if j is exposed or unexposed, respectively, for exposure discordant sets $\mathbf{r}$. This leads to convenient risk weights $W_j = n_0$ or n_1 in (8) for j unexposed or exposed; a sampled subject's relative risk is weighted by the number in that subject's exposure group. It is easy to show that if exposure is the only variable in the model, the counter-matching partial likelihood is proportional to the full cohort partial likelihood. Thus, counter-matching brings the *marginal* full cohort exposure information into the sample.

The general form of the design is defined by the number of "sampling strata" L and the number to be selected from each sampling stratum m_l. The sampled risk set is formed by randomly sampling m_l controls from the n_l subjects in stratum l except for the case's sampling stratum where $m_l - 1$ controls are sampled. The appropriate risk weights to be used in the analysis are n_l/m_l for each subject in sampling stratum l. The choice of L and m_l will depend on the power requirements given assumptions about the information to be collected in the sample and data constraints for the particular problem. We explore these issues in four particular situations where exposure or exposure-related information is available on the full cohort. We also illustrate how to implement the counter-matching design using the Colorado Plateau uranium miners data. This highlights some of the subtleties that arise in applications including counter-matching on a continuous exposure, counter-matching within matching factor levels and the use of time-dependent sampling stratum indicators as well as time-dependent definitions of the sampling strata.

PROBLEM 1.1. A crude exposure surrogate is available on everyone and a more detailed exposure variable is to be collected on a subset. The goal is to assess the effect of the detailed exposure variable.

Suppose that surrogate information Z is available for all cohort members and true measurement information X will be collected on a sample. By calling Z a surrogate measure for X, we mean that Z is correlated to X, but, given X, Z contains no additional information about the rate of disease, that is, the intensity for disease depends only on X through the hazard

$$\lambda(t) = \lambda_0(t)\exp(\beta_X X).$$

For instance, let Z and X be dichotomous and consider sampling a single control for each case. Now, if controls are randomly sampled, case-control pairs are informative only if the control has X status opposite that of the case; if both case and control have identical X, then the partial likelihood contribution from this set is a constant factor (does not depend on β_X) and the pair is noninformative. Intuitively, since Z and X are correlated, selecting controls that have Z status opposite the case should result in more X discordant, that is, informative, pairs than what would have resulted by a simple random sampling of controls. This is precisely what 1:1 counter-matching on the surrogate Z accomplishes. For each case, a control is randomly sampled from those in the risk set with Z value opposite that of the case. True exposure information X is then collected on this counter-matched sample. In the analysis, the risk weights are simply the number of subjects in the risk set with the same Z status as the subject. (While the 1:1 design requires that Z be dichotomous, X is assumed dichotomous for illustration only. There are no restrictions on the form of X nor how it is used in the model.)

Asymptotic relative efficiency results. The improvement in efficiency of counter-matching over random sampling naturally depends upon the accuracy of the surrogate in predicting the true value. Continuing with the 1:1 counter-matching situation, let η be the sensitivity and γ be the specificity of Z for X, and consider the situation where $\beta_X = 0$. The asymptotic variance formulae given in Langholz and Borgan (1995) yields the asymptotic relative efficiency (ARE), relative to 1:1 simple nested case-control sampling, of

$$\text{ARE} = 2[(1-\eta)(1-\gamma) + \eta\gamma]. \tag{16}$$

Further, the full cohort has efficiency 2 relative to the simple 1:1 design. Now if the surrogate is an accurate measure of the true, η and γ will both be close to 1 so that ARE ≈ 2, near full cohort efficiency. Moreover, if the surrogate is a completely "inaccurate" measure of the true in the sense that η and γ are close to zero, we will again approach full cohort efficiency, as discordance in the two surrogate measures, say with values $(0, 1)$, will again in this case lead to discordance in the true measures, however, with values $(1, 0)$.

If, on the other hand, the surrogate and the true measure are "unrelated," in the sense that η and γ are close to 1/2, then ARE ≈ 1, and there is, quite reasonably, no gain in efficiency by counter-matching. In the worst case, Z and X are independent. In this case, $\eta = 1 - \gamma = \text{pr}[Z = 1]$, so that ARE $= 4\eta(1-\eta) \leq 1$ since $0 \leq \eta \leq 1$ and only equals 1 when $\text{pr}[Z = 1] = 1/2$.

While the calculations under $\beta_X = 0$ are enlightening in terms of the general behavior of the relative efficiency of the two sampling designs, realistically, most substudies will be undertaken to investigate exposures that are associated with the disease. Further calculations under the above paradigm show that the efficiency of counter-matching increases with the strength of the association and the rarity of the X (Langholz and Borgan, 1995; Langholz and Clayton, 1994).

Colorado Plateau uranium miners cohort. An issue directly related to counter-matching on a surrogate measure of exposure is how to define sampling

strata (a discrete "surrogate" measure for radon exposure) from the continuous radon exposure variable $R(t)$ that retains as much of the full cohort information about the radon–lung cancer effect as possible. The basic strategy is to form "grouped" radon exposure levels and counter-match based on these, but how to best form these groupings is not obvious. Two solutions to this problem, based on the intuition that it would be desirable to have approximately an equal number of cases in each of the sampling strata, are given in Langholz and Borgan (1995). We illustrate the "empirical distribution" approach here, where the sampling strata are formed using cutpoints at the quantiles of the empirical distribution of *case* radon exposures. Since the distribution of radon exposure of uranium miners changes with age, with older cases tending to have received lower radon exposures than younger cases (see the second and third columns of Table 2), the cutpoints were made to depend on whether the case's age at death was less than or greater than 55. For 1:1 counter-matching, the medians of the exposure distributions were used to form the low and high exposure sampling strata. Thus, from Table 2, a case who was age 53 with 1200 WLM is "low" exposure and is counter-matched to a control randomly sampled from the risk set with "high" exposure, that is, more than 1700 WLM. On the other hand, a case aged 65 with a 1200 WLM exposure is in the "high" exposure sampling stratum and would be counter-matched to a control from the "low" exposure sampling stratum, that is, less than 1000 WLM. This resulted in a sampled data set of 473 distinct subjects, close to the 478 for simple random sampling. For 1:3 counter-matching, the quartiles of the case radon exposure distributions given in Table 2 were used as cutpoints to define four sampling strata. Each case is counter-matched to three controls, one randomly sampled from each sampling stratum other than the case's. This resulted in a sampled set of 765, somewhat less than the 837 in the 1:3 simple data set. This is because the number of subjects in the highest exposure sampling stratum is relatively small and these subjects appear in multiple sampled risk sets more often than would be likely in random sampling (Langholz and Borgan, 1995).

TABLE 2
Quartiles of cumulative radon (WLM) distributions for lung cancer cases and for the risk sets by age group

	Case distribution		Risk set distribution	
	Age < 55	Age ≥ 55	Age < 55	Age ≥ 55
25%	750	500	150	175
50%	1700	1000	400	450
75%	2500	2000	950	1100

For comparison, we used the distribution of all subjects in the risk set as the basis for forming the sampling strata. If this was done for each risk set, one would simply use the empirical distribution of radon exposure at that failure time for all subjects in the risk set. In order to get the appropriate distribution over age groups, we "pooled" the risk sets and formed the empirical distribution of $R(t)$ on these values. The last two columns of Table 2 give the quartiles of this distribution separately for risk sets associated with ages (i.e., the age of the case) up to 55 and those with ages greater than 55. As with the cutpoints based on the case distribution, a 1:1 counter-matched sample was chosen by randomly sampling a control from risk set members on the opposite side of the median of the population distribution. A 1:3 counter-matched sample was based on the quartiles. We note that, while the sampling of controls uses sampling strata defined using cutpoints of the radon exposure distribution, the subject's actual continuous radon exposure values are used in the analysis. As with the simple nested case-control samples (Section 4.4), the counter-matched controls were sampled from risk sets determined by the age and calendar period of death of the case.

The results of the analyses using the sampling strata based on the case distribution of radon exposure are given in the first row, columns 5 and 6, of Table 1 (Univariate model: Radon). Comparing the standard errors of the full cohort, simple nested case-control and counter-matching estimates, it is clear that little radon information is lost in the counter-matched sample relative to the full cohort and that the estimates are substantially more precise than simple nested case-control sampling. The estimates and standard errors from counter-matched samples based on the risk set distributions were 0.45 (0.17) and 0.40 (0.12) for the 1:1 and 1:3 designs, respectively. These are to be compared with the corresponding standard errors (given in Table 1) using the case distribution to form the sampling strata of 0.11 and 0.11. These indicate that basing the cutpoints on the case distribution is the better strategy.

This investigation suggests that sampling strata based on a *surrogate* measure of exposure should be constructed with the goal of splitting the case distribution of the *true* exposure into as evenly divided categories as possible.

PROBLEM 1.2. Exposure data are available on everyone, and confounders are to be collected on a sub-

set. The goal is to assess the effect of exposure after controlling for the confounder.

Let $Z_1(t)$ represent the exposure and $Z_2(t)$ the confounder and consider the intensity model

$$\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1(t) + \beta_2 Z_2(t)\}.$$

We are interested in the precision of $\widehat{\beta}_1$ in this model. Since counter-matching on Z_1 (or a grouped summary of Z_1) brings the marginal information about the exposure into the sample, it is well suited as a sampling design in this situation. Thus, we propose a design where, if necessary, Z_1 is summarized into sampling strata categories and a counter-matched sample is selected based on these sampling strata. The Z_2 confounder information is then collected for the counter-matched sample.

Asymptotic relative efficiency results. Let $Z_1(t)$ and $Z_2(t)$ be dichotomous with a joint distribution, conditional on being at risk, constant over time. In this situation, the asymptotic variance formulas for simple nested case-control sampling and counter-matching are relatively simple and are given in the Appendix of Langholz and Borgan (1995). The asymptotic relative efficiency for estimation of $\widehat{\beta}_1$ of a Z_1 counter-matched sample with a single control (1:1 counter-matched) compared to a simple nested case-control sample with a single control (1:1 simple) is given in the third column of Table 3. Counter-matching clearly results in large gains in efficiency, especially when the relative risk is "away from the null." Also given in the fourth column of Table 3 are the AREs for the estimation of the confounder effect $\widehat{\beta}_2$ after controlling for the exposure Z_1. Apparently, low efficiency for estimation of β_2 is the "price" one pays for high efficiency for the estimation of β_1. Since Z_2 is a confounder, precise estimation of β_2 is less important, so counter-matching has focussed the efficiency to where it is needed.

Colorado Plateau uranium miners cohort. Suppose the goal of our investigation is to assess the effect of radon (available on the full cohort) after adjusting for the confounding effect of smoking (for illustration assumed available only on the counter-matched sample). We used the case distribution based 1:1 and 1:3 radon counter-matched samples of Problem 1.1 and fitted the adjusted model (14). The estimated regression parameters and standard errors from fitting the model for the 1:1 and 1:3 counter-matched samples are given in the fifth and sixth columns ("Adjusted model" rows) of Table 1. Examination of these estimates and standard errors indicates that the counter-matched sample performs well; β_R appears to be substantially better estimated by the counter-matched samples. The difference in precision for the estimation of β_S is not as great as might have been expected based on the relative efficiency results, perhaps due to the commonness of smoking in the cohort and the fact that it is relatively uncorrelated to radon.

TABLE 3

Asymptotic relative efficiencies for the 1:1 *counter-matched design and counter-matching with an additional randomly sampled sampled control* (1:1:1 *hybrid design*) *compared to* 1:1 *simple for* $\mathrm{pr}(Z_1 = 1) = 0.05$, $\mathrm{pr}(Z_2 = 1) = 0.3$ *by the relative risk of* Z_2 (e^{β_2}), *and the odds ratio* $\theta = \mathrm{pr}(Z_1 = 1, Z_2 = 1)\,\mathrm{pr}(Z_1 = 0, Z_2 = 0)/\,\mathrm{pr}(Z_1 = 1, Z_2 = 0)\,\mathrm{pr}(Z_1 = 0, Z_2 = 1)$ *between* Z_1 *and* Z_2

		1:1 counter-matched vs 1:1 simple			1:1:1 hybrid vs 1:1 simple		
e^{β_2}	θ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
			(A) $\exp(\beta_1) = 1$				
0.2	0.2	1.83	0.26	1.26	2.37	1.05	1.35
0.2	1.0	2.00	0.26	1.08	2.28	1.05	1.39
0.2	5.0	1.99	0.25	0.85	2.00	1.07	1.54
1.0	0.2	1.72	0.13	1.43	1.98	1.02	1.87
1.0	1.0	2.00	0.19	1.00	2.00	1.05	1.82
1.0	5.0	1.54	0.20	0.95	1.94	1.05	1.81
5.0	0.2	1.97	0.15	1.49	1.98	1.07	2.65
5.0	1.0	2.00	0.24	1.15	2.48	1.13	2.04
5.0	5.0	1.50	0.28	1.30	2.75	1.15	1.72
			(B) $\exp(\beta_1) = 4$				
0.2	0.2	4.06	0.70	1.90	4.86	1.22	2.02
0.2	1.0	4.20	0.71	1.71	4.60	1.27	2.05
0.2	5.0	3.63	0.70	1.52	3.71	1.35	2.33
1.0	0.2	3.75	0.44	2.92	4.15	1.18	3.59
1.0	1.0	4.35	0.62	2.09	4.35	1.32	3.21
1.0	5.0	3.41	0.65	2.00	3.92	1.34	3.16
5.0	0.2	3.46	0.43	3.27	3.52	1.30	4.86
5.0	1.0	4.07	0.69	2.36	4.62	1.49	3.49
5.0	5.0	3.47	0.80	2.40	5.02	1.55	2.91

Based on the model $\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$. β_3 is the parameter for the interaction term $Z_1 Z_2$.

PROBLEM 1.3. Exposure data are available on everyone, and another covariate is collected on the sample. One goal is to investigate the interaction between the exposure and the other covariate.

Let $Z_1(t)$ represent the exposure, available on all cohort members, and let $Z_2(t)$ be a covariate to be collected on the sample. In addition to exploring the role of Z_2 as a confounder (Problem 1.2), suppose that another goal is to assess the extent to which the relative risk associated with Z_1 varies with Z_2; that is, we are interested in the precision for estimating the interaction parameter β_3 in the model

$$\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1(t) + \beta_2 Z_2(t) + \beta_3 Z_1(t) Z_2(t)\}.$$

Now, if the only goal was to assess interaction, then matching on Z_1 and randomly sampling from the Z_1 homogeneous risk sets is probably the most efficient way to use the full cohort exposure information to sample from the cohort. However, such matching precludes the estimation of β_1 and hence the main effect of Z_1, so that the confounding question could not be addressed and, thus, the matched design is not an acceptable alternative. Of course, the simple nested case-control design is a candidate design. However, since we found in Problem 1.2 that counter-matching is preferable to nested case-control sampling, counter-matching would be a better design here if it is at least as efficient as simple nested case-control sampling for assessing interaction. We now compare the efficiency of these designs for assessing interaction.

Asymptotic relative efficiency results. We consider Z_1 and Z_2 dichotomous and investigate the relative efficiency of $\widehat{\beta}_3$ when $e^{\beta_3} = 1$. The results are given in the fifth column of Table 3, part (A). When $e^{\beta_1} = 1$, that is, when there is no association of Z_1 with disease, the results were mixed. In most of the situations considered, counter-matching was more efficient. However, in two cases, where there was high correlation between Z_1 and Z_2 (odds ratio $\theta = \text{pr}(Z_1 = 1, Z_2 = 1)\,\text{pr}(Z_1 = 0, Z_2 = 0)/\,\text{pr}(Z_1 = 1, Z_2 = 0)\,\text{pr}(Z_1 = 0, Z_2 = 1) = 5.0$), the efficiency of counter-matching for the estimation of β_3 was less than that of nested case-control sampling. In these two situations we examined the proportion of subjects in each of the four Z_1 by Z_2 "cells" and found that counter-matching on Z_1 resulted in a smaller "smallest cell" than random sampling. Thus, we conjecture that, at least when there is no association between Z_1 and disease, the relative efficiency is driven by the smallest cell and that counter-matching will be more efficient if it results in a larger "smallest cell" than random sampling.

When $e^{\beta_1} = 4$, counter-matching results in substantial gains in efficiency, see Table 3, part (B). Thus, for most situations of practical interest, that is, when exposure is associated with disease, these results suggest that counter-matching will be more efficient for the estimation of the interaction parameter than simple nested case-control sampling.

Colorado Plateau uranium miners cohort. Suppose one now wants to investigate whether the effect of radon is smoking dependent; refer to the interaction model parameter estimates in Table 1. It appears that the parameters in the interaction model (15) are somewhat more precisely estimated by the counter-matched samples compared to the simple nested case-control samples, but larger samples would be needed to detect the interaction seen in the full cohort analysis.

PROBLEM 1.4. Exposure 1 data are available on everyone, and exposure 2 data are collected on the sample. The primary goal is to investigate exposure 1, but a secondary goal is to investigate the main effect of the exposure 2.

The exposure counter-matched design when confounder information is collected on the sample is very efficient for estimation of exposure effect while controlling for confounding (Problem 1.2). However, as seen in the fourth column of Table 3, using the model $\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1(t) + \beta_2 Z_2(t)\}$, the efficiency for estimating the effect of the β_2 using the Z_1 counter-matched sample is very low. Since, by definition, one isn't interested in the main effect of confounders, this is acceptable. However, researchers may wish to use the opportunity to assess the role of other potential exposures and collect this "secondary exposure" information for the sampled subjects. Thus, for the purpose of this discussion, we will designate the exposure available on the full cohort as "exposure 1" and the secondary exposure information collected on the sample as "exposure 2." We see that in this terminology, estimation of the main effect of "exposure 2" collected in the counter-matched sample is subject to the same loss of efficiency, relative to simple nested case-control sampling as the confounder in Table 3.

This motivates the compromise *counter-matching with additional randomly sampled controls* design, where we counter-match on the exposure available in the full cohort but pick additional randomly sampled controls. This design, and the risk weights for use in the partial likelihood, are discussed in Borgan, Goldstein and Langholz (1995). For instance, we can consider a cohort where individuals are classified into one of two exposure 1 levels, and use a "1:1:1 design," where we choose a single counter-matched control of opposite exposure 1 level from the case and then randomly sample another control. Intuitively, the 1:1:1 design should have at least the 1:1 counter-matched efficiency for estimating exposure 1 given confounders, and at least the 1:1 nested case-control efficiency for estimating the main effect of exposure 2. This can be shown analytically using the asymptotic relative efficiency formulae given in Borgan, Goldstein and Langholz (1995) and is worked out in Goldstein and Langholz (1995).

Asymptotic relative efficiency comparison. Let the variables Z_1 and Z_2 represent dichotomous

exposures 1 and 2, respectively. The efficiency of the 1:1:1 design relative to 1:1 simple random sampling was calculated under the model $\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1(t) + \beta_2 Z_2(t)\}$. Exposure 1, Z_1, is known for all cohort members and Z_2 is collected on a sample. Analogous to the large sample calculations presented for counter-matching, we computed the relative efficiencies for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in the model given above, that is, the effects of Z_1 and Z_2 after adjusting for each other, and the interaction parameter estimate $\widehat{\beta}_3$ (when $\beta_3 = 0$). The last three columns of Table 3, labeled "1:1:1 Hybrid," reflect the gains anticipated by the above discussion, the hybrid design retains the high 1:1 counter-matching efficiency for estimation of β_1 but also achieves at least the 1:1 simple random sampling efficiency for estimation of β_2. Thus, at the "cost" of an additional control, the 1:1:1 design has the high counter-matching efficiency for investigating exposure 1 and the randomly sampled control "compensates" for the loss of efficiency of counter-matching for investigation of exposure 2 collected in the sample.

Colorado Plateau uranium miners data. While it is not likely that the main effect of smoking would be of interest in a miners study, we give the results of the 1:1:1 design for the uranium miners cohort for illustration. To do this we simply combined the controls from the 1:1 nested case-control sample and the 1:1 counter-matched sample. As expected, the results, given in the last column of Table 1, are a compromise between the two 1:1 samples. The univariate models illustrate this most strikingly with the univariate radon estimate much closer to the 1:1 counter-matched and the univariate smoking estimate much closer to the 1:1 simple. In almost all cases, the estimated 1:1:1 standard errors are smaller than either of those of the 1:1 samples.

PROBLEM 2. *Risk estimation.* The second problem we discuss is one of analysis. We have described how to estimate relative risk parameters β_0 for sampled risk set data and applied these techniques to simple nested case-control and counter-matched samples from the Colorado Plateau uranium miners cohort. It is only recently, however, that methods for estimation of absolute risk have been developed that allow for continuous and time-dependent effects, an important feature of this data. Given a model for the relative risk, the methods outlined in Section 3.4 provide a means to estimate absolute risk of lung cancer and related measures from these samples.

One complication for the uranium miners data is that sampling of controls was restricted to the risk set formed by the case's five-year calendar period of death. Here, it is desirable to "pool" over calendar periods to obtain a single estimate for a given age interval; technical details of this analysis are described in Langholz and Borgan (1996). We computed the predicted risk of lung cancer (10) for a given radon exposure history with constant exposure intensity described by the age at start of exposure, duration of exposure, and total exposure based on model (14). 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. These predicted risks of (with 95% confidence intervals) for various exposure histories during ages 40–49, 50–59 and 60–69 are given in Table 4 from the full cohort, 1:1 simple nested case-control sample and 1:1 counter-matched sample. First, it is clear that radon and smoking vastly increase the risk of lung cancer. The risk of lung cancer in miners who were exposed to 960 WLM over 30 years and were one pack a day smokers (exposures that were quite typical) was between 10 and 20 times the predicted risk in non-radon-exposed nonsmokers. [In spite of the simplicity of the model (14), the predicted risks in non-radon-exposed nonsmokers based on this model are quite close to those actually observed in groups of nonsmokers in the general population (Freedman and Navidi, 1989).]

The sampled data sets produce estimates that are reasonably close to those from the full cohort, with wider confidence intervals than the full cohort, but not that much larger. This suggests that the efficiency of the 1:1 samples for estimation of risk is very good, an observation that has been confirmed by asymptotic relative efficiency calculations (Ørnulf Borgan, personal communication).

PROBLEM 3. *Revisiting a case-control study to collect more precise exposure information.* Suppose that researchers have conducted a nested (or a population-based matched) case-control study and have collected some exposure information on all the subjects in the study. After an analysis of the exposure and other covariate data, the researchers find that it would be desirable to collect additional information on a subset of the case-control subjects, perhaps more precise exposure information or potential confounder information. One could collect the new information for the entire case-control study group or, perhaps, for a random sample of the matched sets. However, intuitively it is advantageous to make use of the exposure information available on the subjects in order to choose an

TABLE 4

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)	Full cohort	1:1 Simple	1:1 Counter-matched
			Age interval 40–49			
—	—	0	0	0.24 (0.13–0.44)	0.16 (0.06–0.42)	0.19 (0.07–0.48)
—	—	0	0.5	0.4 (0.2–0.7)	0.3 (0.2–0.7)	0.4 (0.2–0.8)
20	30	480	0.5	1.0 (0.7–1.4)	0.8 (0.6–1.2)	1.0 (0.7–1.4)
20	30	960	1.0	2.3 (1.7–3.1)	2.0 (1.4–2.9)	2.4 (1.7–3.3)
			Age interval 50–59			
—	—	0	0	0.5 (0.3–1.0)	0.4 (0.2–1.0)	0.4 (0.2–1.1)
—	—	0	0.5	1.1 (0.7–1.8)	1.0 (0.5–2.0)	1.0 (0.5–2.1)
20	30	480	0.5	3.2 (2.5–3.9)	2.9 (2.3–3.8)	3.0 (2.4–3.9)
20	30	960	1.0	7.9 (6.4–9.6)	7.7 (5.7–10.5)	8.0 (6.2–10.4)
			Age interval 60–69			
—	—	0	0	0.7 (0.4–1.3)	0.6 (0.2–1.6)	0.6 (0.2–1.5)
—	—	0	0.5	1.6 (1.0–2.7)	1.7 (0.9–3.5)	1.6 (0.8–3.4)
20	30	480	0.5	4.5 (3.5–5.9)	5.2 (3.9–7.0)	5.0 (3.8–6.6)
20	30	960	1.0	11.7 (9.2–15.0)	14.3 (10.1–20.2)	13.7 (10.1–18.6)

[a]Based on model (14).

[b]Smoking assumed to start at age 20 and continue throughout life.

informative sample. This problem is similar to Problem 1 except that exposure information is only available on a simple nested case-control sample instead of the full cohort. Such a situation could occur in population-based matched case-control studies where investigators wish to collect additional information on study subjects to test new hypotheses, perhaps based on analysis of the case-control study itself or based on hypotheses generated from other sources. We will concentrate on a specific example where investigators wished to make additional exposure measurements on a sample of case-control subjects.

The Swedish case-control study of electromagnetic fields and cancer. In order to investigate the possibility of an association between extremely low-frequency electromagnetic fields (EMF) in the work environment and the development of leukemia and brain tumors, researchers at the National Institute of Occupational Health in Sweden undertook a population-based case-control study (Floderus, Persson, Stenlund, Wennberg and Knave, 1993). The underlying cohort is the male population of mid-Sweden between 1983–1987. Incident cases of leukemia and brain tumors were identified through the Swedish Cancer Registry and, based on physician, patient and close relative permission, were enrolled as a case in the study. Controls were identified using the Swedish Census of 1980 from which, among other information, the gender, date of birth and address of all Swedish residents may be obtained from computerized records. For each case, two controls were randomly sampled from the risk set formed by those who were born in the year of birth of the case and were alive at the time of the study. This resulted in a study group of 250 leukemia and 261 brain tumor cases each matched to two controls. A questionnaire was administered to each subject which asked about factors related to these cancers. For instance, from the job histories, it was determined whether the subject was likely exposured to benzene, solvents generally or ionizing radiation. These were treated as confounders in the analysis. Occupational EMF exposure was estimated for the job performed for the longest period during the 10 years prior to diagnosis of the case. This involved taking gaussmeter measurements at over 1000 workplaces where subjects in this study group were employed. Using standard case-control study analysis methods, the investigators found that there was an increasing trend in risk of leukemia with increasing EMF exposure.

Investigators at the United States National Institute of Occupational Safety and Health (NIOSH) have proposed a biological mechanism of carcinogenesis for EMF radiation and were interested in testing this hypothesis using the Swedish case-

control study group. However, a different EMF measurement, correlated to that used in the original study, is required. These researchers would like to collect additional information on the minimum number of subjects needed to determine, with some certainty, if the new measure is a good predictor of leukemia risk.

Two-stage designs. For our purposes, we consider the original 1:2 matched case-control study as the "first stage" sample from the cohort. The subjects chosen for the NIOSH measurements substudy will be referred to as the "second stage" sample. Further, as in Problem 1.1, we consider the first stage sample EMF measurements Z as a surrogate measure for the second stage true measurement X. Thus, there is no additional information on leukemia risk in Z over that in X and the underlying model is

$$\lambda(t) = \lambda_0(t)\exp(\beta_X X). \tag{17}$$

One obvious second stage sampling design would be to randomly sample matched sets from the original case-control study (design Ia). However, this makes no use of the original EMF measurements. Intuitively, the original EMF measurements could be used to advantage in the second stage sampling since matched sets which have large variation in Z will tend to have large variation in X, that is, be informative sets for assessing the effect of X. We now explore a second stage sampling design (design IIa) that exploits this principle. While there are many measures of variability one might consider, we will use the information (negative second derivative of the log partial likelihood) contribution from the matched set based on an assumed hazard model in Z as this should result in maximizing the inverse of the variance of the estimated parameters. These considerations indicate that the second stage sample consists of the $\tilde{n}$ sets with the largest information contributions based on the model with Z. The number $\tilde{n}$ will be based on the power desired to detect effects of X or dictated by budgetary constraints.

Designs Ia and IIa use both controls in the original matched sets. As a further refinement, we also consider randomly selecting a single control from the two original controls and then sampling the resulting 1:1 sets randomly (design Ib) or according to the Z variability (design IIb). These designs are equivalent to those that start with a first stage sample of 1:1 matched sets. For each of the four designs, the true exposure X would then be obtained for those in the second stage sample.

The analysis of the designs. It is not immediately obvious how to apply our methods to these two-stage designs. In particular, how do we account for the sets that are not sampled? This may be done using a simple trick based on the observation that the partial likelihood contribution (8) from single subject sampled risk sets, that is, those that consist of only the case, is identically 1. Since they contribute nothing to estimation of β_0, they may be dropped from the sample altogether. Thus, for the purpose of developing an estimation method, we may characterize these two-stage procedures by saying that the sample consists of the included sets *plus* sets that consist only of the case (the rejected sets). For the designs we are considering here, $\pi(\mathbf{r}|i)$ is then a distribution over sets $\mathbf{r}$ of size $|\mathbf{r}| = m$ and the singleton set $\{i\}$.

Since each set is chosen with equal probability, it is no surprise that the analysis of designs Ia and Ib are based on the usual unweighted partial likelihood (i.e., $W_i = 1$). Formally, with ρ the probability of sampling a first stage set into the second stage sample, the sampling distribution is given by

$$\pi(\mathbf{r}|i) = \rho\binom{n-1}{m-1}^{-1} I(|\mathbf{r}| = m, i \in \mathbf{r}) + (1-\rho)I(\mathbf{r} = \{i\}),$$

where $|\mathbf{r}|$ is the number of elements in $\mathbf{r}$. So each subject in an included set has the same π which cancels out of the partial likelihood yielding the unweighted partial likelihood.

Not at all obvious, however, is that the partial likelihoods for designs IIa and IIb are also unweighted. To see this, define $V_{\mathbf{r}}(\widehat{\beta}_Z)$ to be the information contribution from the set $\mathbf{r}$ computed using a model for the surrogate measure Z and let $C_i(\kappa)$ be the number of sets $\mathbf{r}$ of size m in the full risk set $\mathscr{R}$ containing i such that

$$V_{\mathbf{r}}(\widehat{\beta}_Z) > \kappa. \tag{18}$$

Then the distribution over sets in the risk set for the design where (1) $m-1$ controls are randomly sampled and (2) the set is included into the final sample if (18) holds is given by

$$\pi(\mathbf{r}|i) = \binom{n-1}{m-1}^{-1} \cdot I(\mathbf{V}_{\mathbf{r}}(\widehat{\beta}_Z) > \kappa,\ i \in \mathbf{r},\ |\mathbf{r}| = m) + \left[1 - C_i(t,\kappa)\binom{n-1}{m-1}^{-1}\right] I(\mathbf{r} = \{i\}). \tag{19}$$

Thus, for included sets, each member of the sampled set has the same value for π yielding the unweighted partial likelihood contribution. The parameter κ is chosen based on the power requirements of the study.

TABLE 5

Power of designs IIa *and* IIb (*picking case-control sets based on* Z *variability*) *for rejecting the null hypothesis* H_0: $\beta_X = 0$, *two sided* $\alpha = 0.05$, *by the number of matched sets and the relative risks per standard deviation in* Z *and* X

		(A) 1:2 matching: Design IIa						
$\exp(\beta_Z)$	$\exp(\beta_X)$	10(30)[a]	15(45)	20(60)	25(75)	30(90)	35(105)	40(120)
1.5	1.8	0.52	0.65	0.75	0.83	0.88	0.92	0.94
	2.0	0.53	0.67	0.78	0.85	0.90	0.94	0.96
	2.5	0.55	0.71	0.82	0.89	0.94	0.96	0.98
1.6	1.8	0.58	0.72	0.82	0.87	0.92	0.95	0.97
	2.0	0.58	0.73	0.83	0.89	0.93	0.96	0.98
	2.5	0.58	0.74	0.84	0.91	0.95	0.97	0.98
1.7	1.8	0.64	0.78	0.87	0.92	0.94	0.97	0.98
	2.0	0.64	0.78	0.87	0.92	0.95	0.97	0.99
	2.5	—	0.78	0.87	0.93	0.96	0.98	0.99
		(B) 1:1 matching: Design IIb						
$\exp(\beta_Z)$	$\exp(\beta_X)$	20(40)	25(50)	30(60)	35(70)	40(80)	45(90)	50(100)
1.5	1.8	0.67	0.75	0.80	0.85	0.88	0.90	0.92
	2.0	0.68	0.76	0.82	0.86	0.89	0.92	0.93
	2.5	0.68	0.77	0.83	0.88	0.91	0.94	0.96
1.6	1.8	0.72	0.80	0.85	0.89	0.92	0.94	0.96
	2.0	0.72	0.80	0.85	0.89	0.92	0.94	0.96
	2.5	0.72	0.80	0.85	0.89	0.93	0.95	0.97
1.7	1.8	0.77	0.85	0.89	0.93	0.94	0.95	0.98
	2.0	0.77	0.85	0.89	0.93	0.94	0.95	0.98

[a]Numbers in parentheses are the total number of subjects in the second stage sample.

We note that in order to choose a κ such that the $\tilde{n}$ most variable sets are selected makes κ data dependent, and this feature is not easily accommodated by the theory. However, we believe that setting κ such that the expected number of selected sets is $\tilde{n}$ in the power and sample size calculation will well approximate the behavior of the actual design.

Power and sample size calculations. The asymptotic theory described in Borgan, Goldstein and Langholz (1995) provides the tools for computing second stage power and sample sizes; the details are given in Goldstein and Langholz (1995). For the EMF measures in the Swedish second stage study, we assume that Z and X are mean and standard deviation normalized and have a joint bivariate normal distribution. With β_Z the limiting value of $\hat{\beta}_Z$ under model (17), one can show that the correlation between Z and X is β_Z/β_X (Xiang and Langholz, 1995). Hence, power and sample size can be parameterized in terms of surrogate and true relative risks. Table 5 gives the power by sample size for the new design for a few e^{β_Z} values near the relative risk observed in the original Swedish study of $e^{\hat{\beta}_Z} = 1.6$ and various e^{β_X}. For comparison, the power when randomly sampling sets (designs Ia and Ib) is given in Table 6. The savings by using the new design can be substantial. For instance, with $e^{\beta_Z} = 1.6$, in order to detect $e^{\beta_X} = 2$ with 90% power, approximately 45 randomly sampled 1:2 matched sets (135 subjects) are required (not shown in table). To achieve the same power, the large Z-variability design (design IIa) requires only 25 sets (75 subjects; Table 5, part (A)). Sampling a single control from each set, then choosing large Z-variabilty sets (design IIb) yields a slight benefit compared to design IIa, but the difference in the total number of second stage subjects needed is small (70 for 1:1 and 75 for 1:2).

6. DISCUSSION

We have illustrated that the model for risk set sampling and the associated analysis methods provide a firm basis for developing designs that are adapted to the goals of a study and the cost of collecting the component pieces of information needed to achieve those goals. Indeed, the conceptual framework elucidated by the model creates an awareness that new designs are possible. We stress, however, that the methods do not provide a way to *generate* an "efficient" design given the goals and circumstances of the study. The counter-matching and other new designs presented here were arrived at after some trial and error. For the time being, developing the appropriate design for a given situation is still an art. The development of new designs in-

TABLE 6
Power of designs Ia *and* Ib (*random sampling of case-control sets*) *for rejecting the null hypothesis* H_0: $\beta_X = 0$, *two sided* $\alpha = 0.05$, *by number of matched sets and relative risk per standard deviation in X*

(A) 1:2 matching: Design Ia							
$\exp(\beta_X)$	10(30)[a]	15(45)	20(60)	25(75)	30(90)	35(105)	40(120)
1.8	0.28	0.39	0.49	0.58	0.66	0.73	0.78
2.0	0.34	0.48	0.59	0.69	0.77	0.83	0.87
2.5	0.46	0.62	0.74	0.83	0.89	0.93	0.96
(B) 1:1 matching: Design Ib							
$\exp(\beta_X)$	20(40)	25(50)	30(60)	35(70)	40(80)	45(90)	50(100)
1.8	0.36	0.44	0.51	0.57	0.62	0.68	0.72
2.0	0.44	0.52	0.60	0.66	0.72	0.77	0.81
2.5	0.55	0.65	0.73	0.79	0.84	0.88	0.91

[a]Numbers in parentheses are the total number of subjects in the second stage sample.

cludes describing putative designs that seem likely to meet the needs of a study in terms of the control sampling distribution $\pi(\mathbf{r}|i)$, and choosing a cost function based on the cost of the component pieces of information and the variance of the parameters of interest. Various designs can then be compared (to each other and to the full cohort design) through the use of the asymptotic variance formulas. Often, insight into features of successful design strategies comes from examination of a variety of these candidate designs. As an example, we consider a design that might be proposed as a solution to Problem 1, where exposure is known for the full cohort and confounder information is to be gathered on a sample. We refer to the terminology and notation used in describing the counter-matching design and, for simplicity, assume that exposure is dichotomized into two sampling strata. The proposed method of sampling is to randomly sample a single control from each sampling stratum. Selecting controls in this way is appealing for the same reason counter-matching is appealing in that there will be increased variability in exposure in the sampled risk set compared to a random sampling of controls. Thus, the sampled risk set will consist of three members, the case (who could be exposed or unexposed) and two controls with opposite exposure status from each other. While at first glance this design may seem quite similar to counter-matching, it is fundamentally different in that the composition of the sampled risk set, in terms of the number of exposed and unexposed, is completely determined by the case's exposure status. Thus, suppose that $\mathbf{r} = \{1, 2, 3\}$ with individuals 1 and 2 exposed and subject 3 unexposed. Now $\pi(\mathbf{r}|1) = \pi(\mathbf{r}|2) = [n_0(n_1 - 1)]^{-1}$, where n_0 and n_1 are the number of unexposed and exposed in the risk set, respectively. However, $\pi(\mathbf{r}|3) = 0$ since subjects 1 and 2 are both exposed and could not have both been picked as controls for 3 under the given sampling scheme. Thus, subject 3's term in the partial likelihood (8) is weighted by zero. Thus, in this situation and generally, only the two subjects from the same sampling stratum contribute to the partial likelihood; clearly not an efficient use of exposure information. The moral of this story is that the structure of the sampled risk set should not "give away" the identity of the case.

While we have focussed on risk set sampling methods and the semi-parametric approach to analysis, we stress that other approaches may be better suited to specific cohort sampling situations. Other design options include the case-cohort (e.g., Kupper, McMichael and Spirtas, 1975; Prentice, 1986; Self and Prentice, 1988) and "grouped time" (e.g., Mantel, 1973; Fears and Brown, 1986; Wild, 1991; Weinberg and Wacholder, 1993) approaches. For instance, if a single set of covariates is to be evaluated for a large number of diseases, a case-cohort sampling design is likely to be a better option than risk set sampling. A grouped time approach may be considered when logistical considerations make control selection on an individual basis difficult. While a discussion of the relative merits and limitations of these approaches is beyond the scope of this paper, we note that risk set sampling is inherently bound to a time scale and any matching factors. Further, because partial likelihood is tied to the proportional hazards model, models that are not of this form currently cannot be easily fitted. [One notable exception is the Aalen linear model (Borgan and Langholz, 1995).] If these are limitations in a particular study, one of the other sampling approaches may be more appropriate.

ACKNOWLEDGMENTS

This work was funded by National Institutes of Cancer Grants CA42949 and CA65123 and National Science Foundation Grants DMS-90-05833 and DMS-95-05075.

REFERENCES

ADELHARDT, M., MØLLER JENSEN, O. and SAND HANSEN, H. (1985). Cancer of the larynx, pharynx and oesophagus in relation to alcohol and tobacco consumption among Danish brewery workers. *Danish Medical Bulletin* **32** 119–123.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1992). *Statistical Models Based on Counting Processes*. Springer, New York.

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

BENICHOU, J. and WACHOLDER, S. (1994). A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine* **13** 651–661.

BOICE, J., BLETTNER, M., KLEINERMAM, R., STOVALL, M. and MOLONEY, W. (1987). Radiation dose and leukemia risk in patients treated for cancer of the cervix. *Journal of the National Cancer Institute* **79** 1295–1311.

BORGAN, Ø., GOLDSTEIN, L. and LANGHOLZ, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23** 1749–1778.

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

BORGAN, Ø. and LANGHOLZ, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model. *Biometrics*. To appear.

BRESLOW, N. and CAIN, K. (1988). Logistic regression for two stage case-control data. *Biometrika* **75** 11–20.

BRESLOW, N. and LANGHOLZ, B. (1987). Nonparametric estimation of relative mortality functions. *Journal of Chronic Diseases* **131** 89S–99S.

BRESLOW, N. and PATTON, J. (1979). Case-control analysis of cohort studies. In *Energy and Health* (N. Breslow and A. Whittemore, eds.), 226–242. SIAM, Philadelphia, PA.

BRESLOW, N. E. and DAY, N. E. (1987). *Statistical Methods in Cancer Research. Volume II. The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

BRESLOW, N. E., LUBIN, J. H., MAREK, P. and LANGHOLZ, B. (1983). Multiplicative models and cohort analysis. *J. Amer. Statist. Assoc.* **78** 1–12.

CLAYTON, D. and HILLS, M. (1993). *Statistical Models in Epidemiology*. Oxford Univ. Press.

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

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

FEARS, T. and BROWN, C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **42** 955–960.

FLODERUS, B., PERSSON, T., STENLUND, C., WENNBERG, Å. Ö. and KNAVE, B. (1993). Occupational exposure to electromagnetic fields in relation to leukemia and brain tumors: A case-control study in Sweden. *Cancer Causes and Control* **4** 465–476.

FREEDMAN, D. and NAVIDI, W. (1989). Multistage models for carcinogenesis. *Environmental Health Perspectives* **81** 169–188.

GARABRANT, D., HELD, J., LANGHOLZ, B. and BERNSTEIN, L. (1988). Mortality of aircraft manufacturing workers in Southern California. *American Journal of Industrial Medicine* **13** 683–693.

GARABRANT, D., HELD, J., LANGHOLZ, B., PETERS, J. and MACK, T. (1992). DDT and related compounds and the risk of pancreatic cancer. *Journal of the National Cancer Institute* **84** 764–771.

GOLDSTEIN, L. and LANGHOLZ, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20** 1903–1928.

GOLDSTEIN, L. and LANGHOLZ, B. (1995). Risk set sampling in epidemiologic cohort studies: Detailed report. Technical Report 101, Dept. of Preventive Medicine, Biostatistics Division, Univ. Southern California, Los Angeles.

HOLFORD, T. (1976). Life tables with concomitant information. *Biometrics* **32** 587–597.

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

KOGEVINAS, M., KAUPPINEN, T., WINKELMANN, R., BECHER, H., BERTAZZI, P., BUENO DE MESQUITA, H., COGGON, D., GREEN, L., JOHNSON, E., LITTORIN, M., LYNGE, E., MARLOW, D., MATHEWS, J., NEUBERGER, M., BENN, T., PANNETT, B., PEARCE, N. and SARACCI, R. (1995). Soft tissue sarcoma and non-Hodgkin's lymphoma in workers exposed to phenoxy herbicides, chlorophenols, and dioxins: Two nested case-control studies. *Epidemiology* **6** 396–402.

KUPPER, L., MCMICHAEL, A. and SPIRTAS, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *J. Amer. Statist. Soc.* **70** 524–528.

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

LANGHOLZ, B. and BORGAN, Ø. (1997). Estimation of absolute risk from nested case-control data. *Biometrics*. To appear.

LANGHOLZ, B. and CLAYTON, D. (1994). Sampling strategies in nested case-control studies. *Environmental Health Perspectives* **102** (Suppl. 8) 47–51.

LIDDELL, F., MCDONALD, J. and THOMAS, D. (1977). Methods of cohort analysis: Appraisal by application to asbestos miners. *J. Roy. Statist. Soc. Ser. A* **140** 469–491.

LUBIN, J. H. and GAIL, M. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics* **40** 63–75.

LUBIN, J., BOICE, J., EDLING, C., HORNUNG, R., HOWE, G., KUNZ, E., KUSIAK, R., MORRISON, H., RADFORD, E., SAMET, J., TIRMARCHE, M., WOODWARD, A., XIANG, Y. and PIERCE, D. (1994). Radon and lung cancer risk: A joint analysis of 11 underground miners studies. NIH Publication 94-3644, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Bethesda, MD.

LUNDIN, F., WAGONER, J. and ARCHER, V. (1971). Radon daughter exposure and respiratory cancer, quantitative and temporal aspects. Joint Monograph 1, U.S. Public Health Service, Washington, DC.

MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29** 479–486.

MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math.* **3** 99–107.

MIETTINEN, O. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* **25** 339–355.

OAKES, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *Internat. Statist. Rev.* **49** 235–264.

PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.

PRENTICE, R. L. and BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65** 153–158.

SELF, S. G. and PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16** 64–81.

THOMAS, D., POGODA, J., LANGHOLZ, B. and MACK, W. (1994). Temporal modifiers of the radon-smoking interaction. *Health Physics* **66** 257–262.

URY, H. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics* **31** 643–649.

WEINBERG, C. and WACHOLDER, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80** 461–465.

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.

WILD, C. (1991). Fitting prospective regression models to case-control data. *Biometrika* **78** 705–717.

XIANG, A. and LANGHOLZ, B. (1995). Comparison of case-control to full cohort analyses when covariates are omitted from the model. Technical Report 108, Dept. of Preventive Medicine, Biostatistics Division, Univ. Southern California, Los Angeles.