# Case–Control Study, Nested

A nested **case–control** study is comprised of subjects sampled from an assembled epidemiological **cohort study** in which the sampling depends on disease status. Nested case–control studies are generally used when disease is rare and, at the minimum, disease outcome has been obtained for all cohort subjects, but it is too expensive to collect and/or process information on **covariates** of interest for the entire cohort. By sampling a small proportion of the nondiseased subjects, there is high cost efficiency for assessing associations between exposures and disease. "Standard" case–control studies, the most common study design in epidemiologic research, may often be viewed as nested case–control studies in which a portion of underlying cohort (usually among the nondiseased) has not been identified [15]. The distinction between standard and nested case–control studies is often ambiguous and, in fact, analysis methods appropriate to standard case–control studies are directly applicable to nested case–control studies. However, depending on the amount of information available in the assembled cohort, there may be a much wider range of design and analysis options for nested case–control studies than for a standard case–control study. So, **confounder** information available in the cohort data is often used to select controls that closely match cases. Also, unlike standard case–control studies, **absolute risk** may often be reliably estimated. Further, it is often possible to compare characteristics of participants to nonparticipants to assess the potential magnitude of selection or information bias (*see* **Bias in Case–Control Studies**). The advantages of nesting a case–control study in a cohort include convenience, cost-efficiency, high validity, and analytic flexibility, for example, [15, 16, 21, 30, 32, 35, 42, 59]. Methodologically, the paradigm of nested case–control sampling is *prospective*, with disease outcome random with probability dependent on covariates. In contrast, the paradigm for standard case–control studies is *retrospective*, with covariates random with distribution depending on disease status. To the extent that standard case–control studies can be viewed as having been sampled from a (perhaps poorly defined) cohort, nested case–control design and analysis developments apply to case–control studies generally.

## Data Model for Nested Case–Control Studies Based on Risk Sets

Cohort data arises by observing a population for disease occurrence over some period of time. So, it is natural to represent nested case–control studies in relation to cohort generation. Figure 1 represents the basic features of a small hypothetical cohort study of 14 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*. 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 follow-up.

The link to nested case–control studies is in the organization of the cohort data into **risk sets** [19]. At any time, the *risk set* is defined to be all subjects under observation. Risk sets may be defined at single points in time, *continuous time risk sets* as in Figure 2 or in time intervals, *grouped time risk sets* as in Figure 3. Risk set members are identified by the "|" at the given time (or time interval). Continuous and grouped risk sets have the structure of individually matched (*see* **Matching**) or unmatched case–control sets, respectively. *Cases* in the risk set are failures at the failure time or time interval, while *controls* are the nonfailures. The nested case–control sample is drawn by sampling from the controls (and possibly from the cases) in the risk sets. Individually matched nested case–control studies arise by sampling from continuous time risk sets at the failure times, while unmatched nested case–control studies arise from sampling from the grouped time risk sets. These are illustrated in Figures 2 and 3, in which ○ represent sampled controls.

## Examples

*Occupational Cohort Study of TCDD Exposure and STS and NHL.* **The International Agency for the Research of Cancer** (IARC) maintains an international register of 21 183 workers exposed to phenoxy herbicides, chlorophenols, and dioxins [52]. In

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

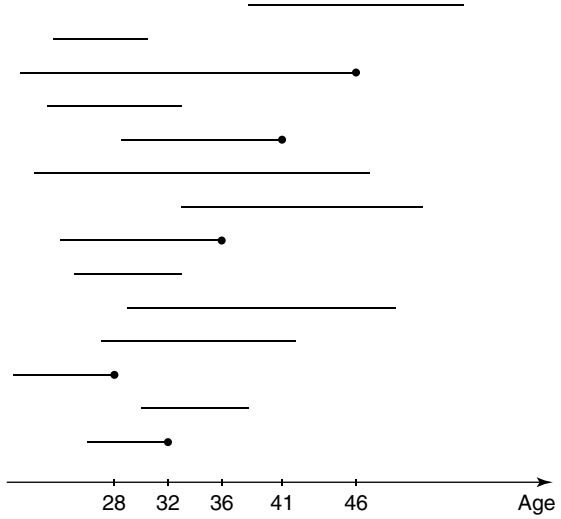


**Figure 1** Cohort of 14 subjects. Each line represents the time on study for one subject. Subjects can either fail (represented by the •) or be censored (no •)

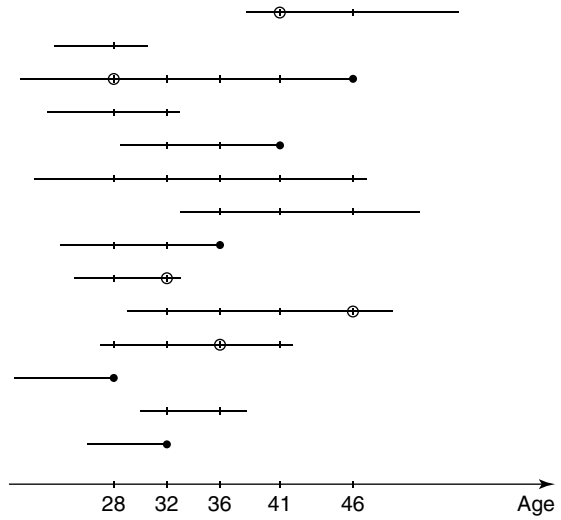


**Figure 2** Continuous time risk sets at each failure time are represented by the "|" marks. The failure is the case in the risk set and the nonfailures are the controls. Single controls, sampled for each case are represented by the ○

a cohort mortality study analysis, standardized mortality ratios (SMRs) (*see*, **Standardization Methods**) of 1.96 and 1.29 were found for soft-tissue sarcomas (STS) and non-Hodgkin's lymphoma (NHL), respectively, comparing exposed to unexposed workers. In

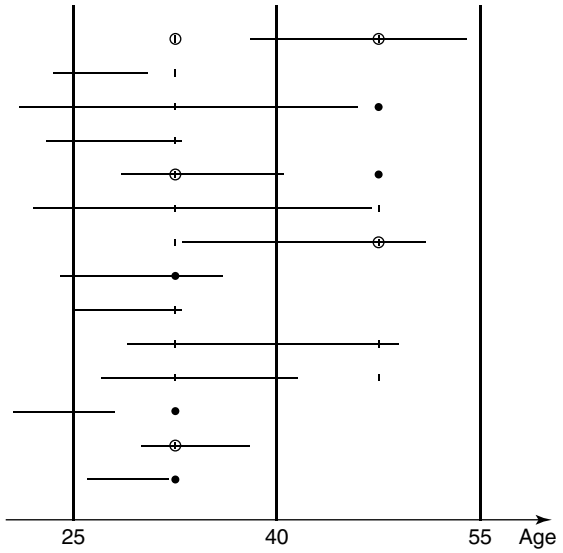


**Figure 3** In this example, grouped time risk sets for 25 to 39 and 40 to 55 age groups are defined as subjects that are on study for any portion of the age interval and are indicated by the "|" marks. Nested case–control sampling in grouped time yields an unmatched case–control study structure with multiple cases per set (the •s) and sampled controls (indicated by ○s). Illustrated here, the number of sampled controls sampled to the number of cases

order to explore the effect of exposure to various agents more fully, a nested case–control study was undertaken in which for each of the 11 STS and 32 NHL cases, five controls were sampled from those from the same country, of the same gender, and same year of birth as the case [27]. For each subject in the nested case–control study, industrial hygienists assessed the degree of exposure to 21 chemicals or mixtures based on company records. Increasing trends of risk of STS and NHL were observed for a number of phenoxy herbicides including 2,4D, and TCDD. This study illustrates a number of potential advantages of nested case–control studies. First, having already assembled the workers cohort, the nested case–control study was a natural follow-up design in order to obtain more detailed exposure information. Second, the workers cohort has much higher **prevalence** of TCDD exposure than general population. Thus, the case–control study selected from this cohort will have much higher statistical **power** to investigate TCDD (and other chemical) associations with STS and NHL than a study of similar size from the general population. Third, on

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

the basis of the $(m-1)/m$ relative efficiency rule [16, 58], this nested case–control study of 258 subjects provides $5/6 = 83\%$ efficiency relative to an analysis of entire cohort of 21 183 for testing associations between single exposures and disease. Finally, because exposure assessment did not require contact with study subjects, recall and selection bias (*see* **Bias in Case–Control Studies**), common problems in standard case–control studies, were avoided.

*Nested Case–control Study of Hypertensive Drugs and the Risk of Myocardio-infarction (MI) within a HMO Cohort*. In this study, the cohort is defined to be patients within the Group Health Cooperative of Puget Sound who were prescribed hypertensive medication for some time during July 1989 through December 1993 [48]. Failures in the cohort were 623 MI cases. Grouped time risk sets were formed on the basis of calendar year and controls were randomly sampled (about 3 times the number of cases) within matching strata based on 10 year age group and gender. For each case–control study member, the types of antihypertensive drugs used were ascertained through computerized records, chart review, and interview. It was found that risk of MI was 60% higher among calcium channel blocker users compared to that among users of either diuretics alone or $\beta$-blockers, a finding that has resulted in a change in treatment strategy. Nesting this case–control study within the HMO cohort had similar advantages to the IARC study. First, the HMO computerized database allowed the identification of cohort members and MI outcome information in a fairly efficient way. There was a high participation rate and, since the type and period of use of drugs could be assessed using the pharmacy database, this information is not subject to information bias.

*Residential Magnetic Field Exposure and Breast Cancer*. The Multiethnic Cohort is a large population-based cohort from Los Angeles and Hawaii of men and women aged 45 to 74 at enrollment between 1993 and 1996. There were 52 112 female Los Angeles County residents who enrolled in the cohort and completed a self-administered questionnaire that included questions about menstrual and reproductive history, use of oral contraceptives and hormone replacement therapy, diet, and physical activity. For the nested case–control study, 751 breast cancer cases diagnosed by 1999 were ascertained through the National Cancer Institute's Surveillance and End Results (SEER) registry in Los Angeles (*see* **Cancer Registries**). Because the study duration was relatively short, the entire study period was considered as a single grouped time risk set. Controls were approximately **frequency matched**, according to the expected number of breast cancer cases, within self-reported ethnicity. Information on traditional breast cancer risk factors was obtained from the cohort baseline questionnaire (100% participation), and each case or control was invited to have an in-home interview about magnetic field exposures (75% participation). Using the baseline residence for questionnaire participants, wire code was obtained for 99% of all case–control subjects, but because permission was required, magnetic field measurements were obtained in homes of only 44% of subjects. No association between magnetic field measures or wire-code and breast cancer were found [36]. Although covariate information obtained through the interview would be subject to the same information biases as a standard case–control study, there was no selection or information bias for the baseline questionnaire and wire-code data. Further, potential bias, in particular, with regard to the missing patterns for magnetic field measurements, could be assessed using the other variables in the baseline questionnaire.

*Nested Case–control Study of the Colorado Plateau Uranium Miners*. The Colorado Plateau uranium miners cohort data were collected to assess the effect of occupational radon exposure on the mortality rates (e.g. [25, 39, 41]) (*see* **Radiation Epidemiology**). 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. Miner radon exposure histories were estimated using job histories and mine radon levels. Although radon and smoking information are available on all cohort members, nested case–control samples with as many as 40 controls per case have been used to reduce the computational burden required to fit complex models exploring the timing of exposures and lung cancer mortality rates [24, 34, 57]. Each of the risk sets was formed by all those who were alive and had entered the study by the

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

age of death of the case and had attained that age in the same five-year calendar period as the case's date of death (matching by calendar time). The analyses based on the nested case–control data closely approximates the corresponding **Cox regression**, but fitting the models required a fraction of computing time; in the case of the latency models, this meant a reduction from a few days to less than an hour. Further, in the nested case–control data, radon and smoking summaries need only be computed at a fixed (failure) time rather than dynamically in a Cox regression. This makes it much easier to identify data errors and check exposure calculation routines. Absolute risk of lung cancer death, given radon and smoking histories, were estimated from nested case–control data from this cohort [29].

## Statistical Analysis Based on the Proportional Hazards and Odds Models for Sampled Risk Set Data

Any of the analysis methods available for "standard" case–control studies, including **conditional** and unconditional **logistic regression** and **Mantel–Haenszel** methods may be used to estimate rate or **odds ratios** from a nested case–control study when the sampling is "simple". Here, we describe methods that are based on the risk set sampling data model that accommodates quite general sampling.

### *Proportional Hazards and Odds Models*

The standard methods for analysis of nested case–control data correspond to and are generalizations of estimation methods for cohort data based on risk sets. Data analysis methods are derived from **semiparametric** models for disease occurrence, the **proportional hazards** or **proportional odds models** being appropriate to continuous or grouped time data, respectively [19]. Each is assumed to have **multiplicative** form

$$\lambda(t; z(t)) = \lambda_0(t) r(z(t); \beta_0) \tag{1}$$

where $r(z(t); \beta)$ is the rate (odds) ratio of disease for an individual with covariates $z(t)$ at time $t$ and $r(0; \beta) = 1$, so $\lambda_0(t)$ is the rate (odds) of disease in subjects with $z = 0$. In continuous time, $t$ refers to any time, while in grouped time the $t$ is discrete and indexes the time intervals. The proportional hazards model may be obtained as the limit to the proportional odds model as the time interval lengths go to zero. As a consequence, the rate ratio parameter and odds ratio parameter in grouped time structure will be close when the probability of failure (rare disease) in each time interval is small.

### *Estimation of Rate Ratio Parameters from Continuous Time Data*

The **partial likelihood** method for cohort data is based on the probability that a subject is a case given the risk set [19, 20]. Similarly, the partial likelihood for nested case–control data is based on the probability that a subject is a case given the case–control set [6, 35, 43, 55]. This will depend on the sampling method and leads to a **likelihood** of the form

$$\prod_{\text{failure times}} \frac{r_{\text{case}}(\beta)\pi_{\text{case}}}{\sum_{k\in\tilde{\mathcal{R}}} r_k(\beta)\pi_k} \tag{2}$$

where $\tilde{\mathcal{R}}$ is the case–control set, the $r_k$ are computed at the failure time, and $\pi_k$ is the probability of picking the particular case–control set if $k$ was the case. These will generally be replaced by a convenient weights $w_k$ that are proportional to the $\pi_k$.

For instance, for **simple random sampling** of $m-1$ controls from the $n-1$ in the risk set, $\pi_k = \binom{n-1}{m-1}^{-1}$. In this case, the $\pi_k$ are the same for all case–control set members so we may take $w_k = 1$, which yields the "unweighted" conditional likelihood for standard matched case–control data (*see* **Matched Analysis**). Standard conditional logistic regression software may be used to estimate the rate ratio ($\beta_0$) parameters (*see* **Software, Biostatistical**).

### *Estimation of Odds Ratio Parameters from Grouped Time Data*

Parallel to the continuous time situation, a partial likelihood is based on the probability that a set of subjects **D** are the cases given that the case–control set is $\tilde{\mathcal{R}}$ and is given by

$$\prod_{\text{grouped times}} \frac{\lambda^{|\mathbf{D}|}\, r_{\mathbf{D}}(\beta)\pi_{\mathbf{D}}}{\sum_{\mathbf{s}\subset\tilde{\mathcal{R}}} \lambda^{|\mathbf{s}|} r_{\mathbf{s}}(\beta)\pi_{\mathbf{s}}}, \tag{3}$$

where $r_{\mathbf{s}}(\beta) = \prod_{j\in\mathbf{s}} r(Z_j; \beta)$, $|\mathbf{s}|$ is the number of elements in $\mathbf{s}$, and $\pi_{\mathbf{s}}$ is the probability of picking the case–control set given that $\mathbf{s}$ is the set of

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

cases. For analysis, the $\pi_{\mathbf{s}}$ can be replaced by convenient weights $w_{\mathbf{s}}$ that are proportional to $\pi_{\mathbf{s}}$ [31]. For instance, in $1:m-1$ frequency matching, the number of controls randomly sampled is $m-1$ times the number of cases. Then, the $\pi_{\mathbf{s}} = \binom{n-|\mathbf{D}|}{m-|\mathbf{D}|}^{-1}$ for all subsets $\mathbf{s}$ of the case–control set $\tilde{\mathcal{R}}$ that are of the same size as the case set. Cancellation of the common $\pi$ from numerator and denominator leads to the standard (unweighted) conditional logistic likelihood for unmatched case–control data. On the other hand, case–based sampling (for example, [28]) in which a random sample of $m|\mathbf{D}|$ subjects (without regard to failure status) is drawn from the cohort and additional failures are included is "weighted" with $w_{\mathbf{s}} = \binom{|\mathbf{s}|}{m|\mathbf{D}|-(|\tilde{\mathcal{R}}|-|\mathbf{s}|)}$ [31].

*Conditional Logistic Likelihood.* A likelihood estimator that is closely related to the partial likelihood conditions on the number of cases so that the *conditional logistic likelihood* is given by

$$\prod_{\text{grouped times}} \frac{r_{\mathbf{D}}(\beta)\pi_{\mathbf{D}}}{\sum_{\mathbf{s}\subset\tilde{\mathcal{R}}:|\mathbf{s}|=|\mathbf{D}|} r_{\mathbf{s}}(\beta)\ \pi_{\mathbf{s}}}. \tag{4}$$

Unlike the partial likelihood (3) from which the baseline odds may often be estimated, the baseline odds parameter is conditioned out of (4). Also, unlike the partial likelihood, for each of the standard (simple) control selection methods, including frequency matching, Bernoulli trials (*see* **Binary Data**) and case-base, the conditional likelihood is the same, "unweighted" version [3]. The conditional likelihood is often used when the number of cases and/or controls in all or some case–control sets is small or when there are tied failure times in continuous time analyses (*see* **Tied Survival Times**).

*Unconditional Logistic Regression.* This is the most commonly used alternative for analysis of grouped time nested case–control studies with random sampling and is based on the product of "marginal" case/control probabilities within the case–control set

$$\prod_{\text{grouped times}} \prod_{j\in\tilde{\mathcal{R}}} \frac{[\lambda\ w_j\ r_j(\beta)]^{D_j}}{1+\lambda\ w_j\ r_j(\beta)}, \tag{5}$$

where $D_j$ is a case–control status indicator, $w_j$ is a marginal "inverse control sampling probability". The $w_j$ depend both on $\beta_0$ and $j$, but for common situations, this dependence is small for "large samples". Further, because the probabilities are only marginal, the variance cannot generally be estimated as the "inverse information" (e.g., [11, 31]). The unconditional logistic likelihood also arises from two-phase studies, closely related to nested case–control studies, in which the cohort is taken as a fixed set of cases and controls from which a second stage sample is drawn from the first, for example, [11, 13, 53, 62, 67] (*see* **Case–Control Study, Two-phase**). In the case of simple random sampling (with $m-1$ controls per case), the $w_j$ are approximately all equal to $(n-|\mathbf{D}|)/(m|\mathbf{D}|-|\mathbf{D}|)$ and a **nuisance parameter** $\theta_0$ can be used in place of $\lambda_0\ w_j$ in the likelihood. In this special case, the variance of the odds ratio parameter may be estimated using the standard inverse information estimator [3].

### *Absolute Risk Estimation*

Unlike a standard case–control study in which the cohort cannot be identified from a nested case–control study, it is possible to estimate the baseline **hazard rate** and, more generally, absolute risk quantities that are functions of the hazard.

*Estimation of Risk from Continuous Time Data.* If the number at risk $n$ and the control selection probabilities $\pi_k$ are known at each failure time, then **cumulative hazard** functions and survival functions may be estimated using a generalization of the Breslow estimator of the baseline hazard [1, 19, 29] (*see* **Hazard Ratio Estimator**). Let $z^0(t)$ be a covariate history and $r^0(t) = r(z^0(t);\beta_0)$ be the rate ratio (as a function of time) associated with $z^0$ according to the model. The basic components for estimators of risk are the jumps in the hazard at the failure times. With $n$ the number at risk and $\hat{r}_k = r(z_k, \hat{\beta})$, the relative risk for individual $k$ predicted using $\hat{\beta}$, the hazard jump from a case–control set is estimated by

$$\hat{r}^0 \Big/ n \sum_j \frac{\pi_j}{\sum_k \pi_k}\hat{r}_j, \tag{6}$$

where the sums are over case–control set members [29]. Note that setting $m=n$ and $z^0(t)\equiv 0$ yields the Breslow estimator of the baseline hazard for the full cohort [19]. For simple random sampling of $m-1$ controls, $\pi_j/\sum_k \pi_k = m$ so the denominator is given by $n/m\sum \hat{r}_k$. Cumulative hazard

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

and **Kaplan–Meier-type estimators** of risk and an **Aalen–Johansen-type estimator** of risk in the presence of **competing** causes of failure, as well as corresponding variance estimators have been described with application to the Colorado Plateau uranium miners study [9, 29].

*Estimation of Risk from Grouped Time Data.* Estimation of absolute risk from grouped time nested case–control data with general sampling is a topic of continuing research. When the overall risk of disease is known in the cohort, one approach uses the distribution of covariates in controls as representative of the cohort rates to infer risk within exposure subgroups [4]. When the $w_j$ can be specified, then the baseline odds (and hence the risk) can evidently be estimated using the grouped time partial (3) or the unconditional (5) likelihoods.

### *Asymptotic Properties and Efficiency*

The "likelihoods" for both continuous and grouped time are "partial" in the same sense as the Cox partial likelihood for full cohort data [19] in that the same subject may appear in multiple sets. In an extension to the **counting process** and martingale theory approach for full cohort data [2], nested case–control data is represented by a counting process $N_{i,\mathbf{r}}(t)$ for occurrences of both subject $i$ becoming diseased and $\mathbf{r}$ the case–control set. Within this framework, the case–control set variability is constant, with sample size and the asymptotics driven by the increasing number of case–control sets. Conditions for the consistency and asymptotic normality of the partial likelihood rate ratio and baseline hazard estimators have been described for a wide range of sampling methods [6, 23]. Also provided are expressions for the asymptotic variance from which **efficiency**, statistical **power**, and **sample size** calculations can be made. For simple sampling, these are refinements of those for standard individually matched case–control studies that take into account the underlying failure time structure [23]. Performance of the partial likelihood under model **misspecification** under simple sampling has also been studied [63, 64]. For grouped time case–control data, the framework can be similarly defined with $N_{\mathbf{d},\mathbf{r}}(t)$ now indicating the set $\mathbf{d}$ of diseased subjects and $\mathbf{r}$ the case–control set. However, the asymptotic theory will depend on if and how the time intervals "shrink" as a function of sample size. For fixed time intervals, the number of case–control sets is fixed and the asymptotics are driven by increasing sample size within case–control sets and thus, the asymptotic theory is very different than in the continuous time situation. The theory for the "unweighted" conditional logistic (4) and unconditional logistic (5) likelihoods based on rejective sampling has been described [3]. However, a general theory has not been derived in the grouped time setting.

### *Other Approaches to Estimation and Other Models*

*Proportional Hazards Models.* **Mantel–Haenszel** estimators for nested case–control studies with simple random sampling have been described and shown to be consistent [65, 66]. Methods for estimation of relative mortality have been described on the basis of an extension of the proportional hazards model [7, 14, 54].

Another class of estimators seeks to use case and control information at times other than when they were sampled, all with the goal of capturing more information from the case–control sample than the partial likelihood. There have been a number of methods that enlarge or restrict the controls used in the unweighted version of (2) in order to increase efficiency [33, 45, 49, 60]. Interestingly, even though these methods made "better" use of the sample, it was found that efficiency gains were modest at best, and often were worse in situations of practical importance [33]. Another method incorporates external rates and estimation is based on joint cohort Poisson and nested case–control partial likelihoods [56]. Methods have been proposed using an "inverse weighting" method [18, 51] as well as an estimator based on "local averaging". The latter was shown to be more efficient than earlier extensions and is more efficient than the partial likelihood in a range of situations [17]. All these methods show some improvement when disease is common and/or the rate ratio is large. Further work is necessary to establish the practical guidelines for when these methods offer significant benefits over the standard methods.

*Nonproportional Hazards Models and Extensions.* Methods for modeling **excess risk** and estimation of absolute risk based on the Aalen linear regression

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

model, from nested case–control studies have also been developed [8]. Nested case–control studies with appropriately sampled controls can be used to estimate transition rate ratio parameters for recurrences or multiple outcomes [37, 38] and for **Markov** transition probabilities [5]. A useful method for estimation of parameters in parametric models has been described [51].

### *Control Sampling Methods*

*General Guidelines*. The likelihood methods are all valid (and most useful) if the risk sets are sampled independently over time so that subjects may serve as controls in multiple case–control sets and failures may be controls in "earlier" risk sets. Restricting controls to be used only once or using "pure" controls, those that never become cases, will, in theory, result in biased estimation [40] unless special analysis methods are used [45]. However, if disease is rare, this bias is generally negligible. Very general sampling methods can be accommodated by the risk set sampling likelihoods but a general guideline for useful designs is that the structure of the case–control set should not reveal the identity of the case [30].

*Matching and Random Sampling*. The most commonly used methods for selecting controls is to randomly sample from risk set members who "match" on a set of factors. For continuous time case–control studies, this means that the sampled controls will be similar to the individual case on these factors. For grouped time studies, the risk set will be partitioned into matching strata and controls (and cases) are sampled from within these strata. Although the choice of matching criteria will depend on the needs of the study, common matching factors include gender, race/ethnicity, calendar year, and/or year of birth. The latter is often desirable because, as in the hypertensive drug-MI example, a natural timescale is age but matching on year of birth aligns the cases and controls with respect to calendar time and thus assures comparable data quality and control for "secular trends" in diagnostic treatment practices [12].

*Fine Matching*. When a continuous matching factor is available on all cohort members, nearest neighbor and caliper matching are possible as continuous time control selection options [26]. For instance, in a study of occupational exposure to chemical agents and pancreatic cancer, for each case, the four controls in the risk set who most closely matched on date of birth were enrolled into the nested case–control study [22]. In this study, there was no random sampling at all and the nearest neighbor matching completely determined the control selection. The unweighted partial likelihood is not strictly correct under this sampling, but conditions have been described when it is asymptotically valid, as well as other analysis options when the matching factor is included as a covariate in the proportional hazard model [26].

*Exposure Stratified Sampling Methods*. Until recently, it was thought that control selection could not depend on exposure related variables [50]. In fact, unbiased estimation is possible if the appropriate control selection probabilities (weights) are specified in the likelihood. Such designs include **counter-matching**, variants of quota sampling, two-stage sampling, and exposure stratified case-base (**case–cohort**) sampling [6, 10, 11, 30, 31, 61].

### *Issues Related to Grouped Time Studies*

*Continuous Time as a Limiting Case of the Grouped Time Model*. Parallel to the approach of Cox for cohort data organized into risk sets, individually matched case–control study designs and methods can be obtained from unmatched studies by "shrinking the time interval" to zero. Thus, the proportional odds model converges to the proportional hazards, $1 : m - 1$ frequency matching becomes $1 : m - 1$ individual matching and the grouped time partial (or conditional) likelihood converges to the continuous time partial likelihood.

*Grouped Time as an Approximation to Reality*. Because time is in reality "continuous", the grouped time approach necessarily involves a number of approximations, which may be critical when the grouped time intervals are large. One issue is the ambiguity in the definition of who is in the risk set, in particular, among those who are censored during a grouped time interval. The problems that can arise correspond to those associated with ignoring censoring in failure time data (but on the scale of the time interval). In Figure 3, we have defined subjects as being at risk in the interval if they are at risk during any part of the interval. Another problem with

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

grouping time is that there is not a single unambiguous "reference time" from which to compute time-dependent covariates. Various strategies have been to use the average time for the cases, as was done in the residential magnetic fields breast cancer example [36] or to randomly assign times to the controls based on the case failure time distribution within the interval. Unless there are large changes in at-risk status and/or covariate values over time intervals, strategies that reasonably approximate the continuous time risk sets will yield estimates that are close to the corresponding continuous time estimator.

*Failure Time Analysis of Grouped Time Case–control Studies*. A number of methods have been developed to estimate rate ratio parameters when the case–control study is sampled from grouped time risk sets. Most notable of these is the case–cohort method [44]. To see that this is a grouped time sampling, note that the sampling is the "case-base" method [28] in which a random sample of subjects (without regard to failure status) is drawn from the cohort and additional failures are included. The grouped time analysis methods based on estimation of *odds ratios* for exposures apply with the time interval taken as the entire study period; this analysis is subject to the pitfalls associated with grouping time described above. The case–cohort *analysis* method allows estimation of the *rate ratios* appropriately accounting for censoring and **time-dependent covariates** from the case-based sampled data. This idea was generalized to estimators of rate ratios from "simple" unmatched case–control data from the cohort [17]. A comparison of nested case–control and case–cohort approaches is given in the article **Case–Cohort Study**.

## *Nested and Standard Case–control Studies*

*Relevance of Nested Case–control Studies*. Standard case–control studies may often be viewed as a nested case–control study within a nonassembled (and perhaps poorly defined) cohort. Thus, methods developed for nested case–control studies that do not require further knowledge of cohort information will apply to standard case–control designs. So, a number of study designs, including quota sampling, modified randomized recruitment, and individually matched two-stage studies have been proposed on the basis of the nested case–control study paradigm that do not require an assembled cohort [6, 30, 31]. Another example is a robust (*see* **Robustness**) variance estimator derived for continuous time $1 : m - 1$ nested case–control studies that may be used in standard individually matched case–control study analysis [64].

*Difference in Methodological Approach for Nested and Standard Case–control Studies*. Methodologically, perhaps the biggest difference between nested and "nonnested" case–control studies is the data model used to develop methods. Traditionally, case–control data is viewed as generated "retrospectively", with individual exposure independent random quantities, with distribution conditional on disease status (e.g. [46]). A key result is that, even under the retrospective model, odds ratio parameters are estimable using the unconditional logistic likelihood (5), (e.g. [47]). For simple sampling, either approach leads to similar methods and inference about odds ratio parameters in grouped time data [3]. However, methods developed under the risk set sampling model provide a connection to failure time cohort data and associated methods and, further, provide a natural framework for the development of methods for individually matched case–control data.

## *References*


[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

[2] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[3] Arratia, R., Goldstein, L. & Langholz, B. (2005). Local central limit theorems, the high order correlations of rejective sampling, and logistic likelihood asymptotics, *Annals of Statistics*; to appear.

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

[5] Borgan, Ø (2002). Estimation of covariate-dependent markov transition probabilities from nested case-control data, *Statistical Methods in Medical Research* **11**, 183–202.

[6] Borgan, Ø, Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.

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


*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

[8] Borgan, Ø & Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model, *Biometrics* **53**, 690–697.

[9] Borgan, Ø & Langholz, B. (1998). Risk set sampling designs for proportional hazards models, in *Statistical Analysis of Medical Data: New Developments*, B.S. Everitt & G. Dunn, eds. Arnold, London, pp. 75–100.

[10] Borgan, Ø, Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.

[11] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two stage case-control data, *Biometrika* **75**, 11–20.

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

[13] Breslow, N.E. & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood, and maximum likelihood methods for logistic regression analysis of two-stage data, *Statistics in Medicine* **16**, 103–116.

[14] Breslow, N.E. & Langholz, B. (1987). Nonparametric estimation of relative mortality functions, *Journal of Chronic Diseases* **131**(Suppl. 2), 89S–99S.

[15] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[16] Breslow, N.E. & Patton, J. (1979). Case-control analysis of cohort studies, in *Energy and Health*, N.E. Breslow & A.S. Whittemore, eds. SIAM Institute for Mathematics and Society, SIAM, Philadelphia, pp. 226–242.

[17] Chen, K. (2001). Generalized case-cohort sampling, *Journal of the Royal Statistical Society, Series B, Methodological* **63**, 791–809.

[18] Chen, K. & Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model, *Biometrika* **86**, 755–764.

[19] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **34**, 187–220.

[20] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[21] Ernster, V.L. (1994). Nested case-control studies, *Preventive Medicine* **23**, 587–590.

[22] Garabrant, D.H., Held, J., Langholz, B., Peters, J.M. & Mack, T.M. (1992). DDT and related compounds and the risk of pancreatic cancer, *Journal of the National Cancer Institute* **84**, 764–771.

[23] Goldstein, L. & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model, *Annals of Statistics* **20**, 1903–1928.

[24] Hauptmann, M., Behrane, K., Langholz, B. & Lubin, J.H. (2001). Using splines to analyze latency in the colorado plateau uranium miners cohort, *Journal of Epidemiology and Biostatistics* **6**, 417–424.

[25] Hornung, R.W. & Meinhardt, T.J. (1987). Quantitative risk assessment of lung cancer in U. S. uranium miners, *Health Physics* **52**, 417–430.

[26] Kim, S. & De Gruttola, V. (1999). Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial, *Lifetime Data Analysis* **5**(2), 149–72.

[27] Kogevinas, M., Kauppinen, T., Winkelmann, R., Becher, H., Bertazzi, P.A., Bueno de Mesquita, H.B., Coggon, D., Green, L., Johnson, E., Littorin, M., Lynge, E., Marlow, D.A., Mathews, J.D., Neuberger, M., Benn, T., Pannett, B., Pearce, N. & 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.

[28] Kupper, L.L., McMichael, A.J. & Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk, *Journal of the American Statistical Association* **70**, 524–528.

[29] Langholz, B. & Borgan, Ø (1997). Estimation of absolute risk from nested case-control data, *Biometrics* **53**, 767–774.

[30] Langholz, B. & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies, *Statistical Science* **11**, 35–53.

[31] Langholz, B. & Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling, *Biostatistics* **2**, 63–84.

[32] Langholz, B., Rothman, N., Wacholder, S. & Thomas, D.C. (1999). Cohort studies for characterizing measured genes, *Monographs Journal of the National Cancer Institute* **26**, 39–42.

[33] Langholz, B. & Thomas, D.C. (1991). Efficiency of cohort sampling designs: some surprising results, *Biometrics* **47**, 1563–1571.

[34] Langholz, B., Thomas, D.C., Xiang, A. & Stram, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the colorado plateau uranium miners cohort, *American Journal of Industrial Medicine* **35**, 246–256.

[35] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos miners, *Journal of the Royal Statistical Society A* **140**, 469–491.

[36] London, S.J., Pogoda, J.M., Hwang, K.L., Langholz, B., Monroe, K.R., Kolonel, L.N., Kaune, W.T., Peters, J.M. & Henderson, B.E. (2003). Residential magnetic field exposure and breast cancer risk in the multiethnic cohort study, *American Journal of Epidemiology* **158**, 969–980.

[37] Lubin, J.H. (1985). Case-control methods in the presence of multiple failure times and competing risks, *Biometrics* **41**, 49–54.

[38] Lubin, J.H. (1986). Extensions of analytic methods for nested and population-based incident case-control studies, *Journal of Chronic Diseases* **39**, 379–88.

[39] Lubin, J.H., Boice, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A., Xiang, Y.S. & Pierce, D.A. (1994). *Radon and Lung*


*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025

*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.

[40] Lubin, J.H. & Gail, M.H. (1984). Biased selection of controls for case-control analyses of cohort studies, *Biometrics* **40**, 63–75.

[41] Lundin, F.D., Wagoner, J.K. & Archer, V.E. (1971). *Radon daughter exposure and respiratory cancer, quantitative and temporal aspects*. Joint Monograph 1, U.S. Public Health Service, Washington.

[42] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.

[43] Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion), *International Statistical Review* **49**, 235–264.

[44] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[45] Prentice, R.L. (1986). On the design of synthetic case-control studies, *Biometrics* **42**, 301–310.

[46] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[47] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**, 403–411.

[48] Psaty, B.M., Heckbert, S.R., Koepsell, T.D., Siscovick, D.S., Raghunathan, T.E., Weiss, N.S., Rosendaal, F.R., Lemaitre, R.N., Smith, N.L., Wahl, P.W., et al. (1995). The risk of myocardial infarction associated with antihypertensive drug therapies, *Journal of the American Medical Association* **274**, 620–625.

[49] Robins, J.M., Prentice, R.L. & Blevins, D. (1989). Designs for synthetic case-control studies in open cohorts, *Biometrics* **45**, 1103–1116.

[50] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott-Raven Publishers, Philadelphia.

[51] Samuelsen, S.O. (1997). A pseudolikelihood approach to analysis of nested case-control data, *Biometrika* **84**, 379–394.

[52] Saracci, R., Kogevinas, M., Bertazzi, P.A., Bueno de Mesquita, B.H., Coggon, D., Green, L.M., Kauppinen, T., L'Abbé, K.A., Littorin, M., Lynge, E., Mathews, J.D., Neuberger, M., Osman, J., Pearce, N. & Winkelmann, R. (1991). Cancer morality in workers exposed to chlorophenoxy herbicides and chlorophenols, *Lancet* **338**, 1027–1032.

[53] Scott, A.J. & Wild, C.J. (1991). Fitting logistic models under case-control or choice based sampling, *Journal of the Royal Statistical Society Series B* **48**, 170–182.

[54] Suissa, S., Edwardes, M.D. & Boivin, J.F. (1998). External comparisons from nested case-control designs, *Epidemiology* **9**, 72–78.

[55] Thomas, D.C. (1981). General relative-risk models for survival time and matched case-control analysis, *Biometrics* **37**, 673–686.

[56] Thomas, D.C., Blettner, M. & Day, N.E. (1992). Use of external rates in nested case-control studies with application to the international radiation study of cervical cancer patients, *Biometrics* **48**, 781–794.

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

[58] Ury, H.K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, *Biometrics* **31**, 643–649.

[59] Wacholder, S. (1995). Design issues in case-control studies, *Statistical Methods in Medical Research* **4**, 293–309.

[60] Wacholder, S., Gail, M.H. & Pee, D. (1991). Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort, *Biometrics* **47**, 63–76.

[61] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling, *Biometrics* **46**, 963–975.

[62] Whittemore, A. (1997). Multistage sampling designs and estimating equations, *Journal of the Royal Statistical Society B* **59**, 589–602.

[63] Xiang, A.H. & Langholz, B. (1999). Comparison of case-control to full cohort analyses under model misspecification, *Biometrika* **86**, 221–226.

[64] Xiang, A.H. & Langholz, B. (2003). Robust variance estimation for rate ratio parameter estimates from individually matched case-control data, *Biometrika* **90**, 741–746.

[65] Zhang, Z.-Z. (2000). On consistency of Mantel-Haenszel type estimators in nested case-control studies, *Journal of the Japan Statistical Society* **30**(2), 205–211.

[66] Zhang, Z.-Z., Fujii, Y. & Yanagawa, T. (2000). On Mantel-Haenszel type estimators in simple nested case–control studies, *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]* **29**, 2507–2521.

[67] Zhao, L.P. & Lipsitz, S. (1992). Designs and analysis of two-stage studies, *Statistics in Medicine* **11**, 769–782.

BRYAN LANGHOLZ

*Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03025