Case–Cohort Study

The case–cohort design is a method of sampling from an assembled epidemiologic **cohort study** or **clinical trial** in which a **random sample** of the cohort, called the *subcohort*, is used as a comparison group for all cases that occur in the cohort. This design is generally used when such a cohort can be followed for disease outcomes but it is too expensive to collect and process **covariate** information on all study subjects. Though it may be used in other settings, it is especially advantageous for studies in which covariate information collected at entry to the study is "banked" for the entire cohort but is expensive to retrieve or process (see examples below) and multiple disease stages or outcomes are of interest. In such circumstances, the work of covariate processing for subcohort members can proceed at the beginning of the study. As time passes and cases of disease occur, information for these cases can be processed in batches. Since the subcohort data are prepared early on and are not dependent on the occurrence of cases, statistical analyses can proceed at regular intervals after the processing of the cases. Furthermore, staffing needs are quite predictable. Motivated by the case–base sampling method for simple **binary** outcome data [15, 23], Prentice described the design and a **pseudo-likelihood** method of analysis (see below) for the case–cohort design.

Design

The basic components of a case–cohort study are the *subcohort*, a sample of subjects in the cohort, and *nonsubcohort cases*, subjects that have had an event and are not included in the subcohort. The subcohort provides information on the **person-time** experience of a random sample of subjects from the cohort or random samples from within strata (*see* **Stratification**) of a **confounding** factor. In the latter situation, differing sampling fractions could be used to align better the person-time distribution of the subcohort with that of the cases. Methods for sampling the subcohort include sampling a fixed number without replacement [26] (*see* **Sampling With and Without Replacement**) and sampling based on independent Bernoulli "coin flips" [34] (*see* **Binomial Distribution**). The latter may be advantageous when subjects are entered into the study prospectively; the subcohort may then be formed concurrently rather than waiting until accrual into the cohort has ended [30, 34]. Simple case–cohort studies are the same as case–base studies for simple **binary** outcome data. But, in general, portions of a subject's time on study might be sampled. For example, the subcohort might be "refreshed" by sampling from those remaining on study after a period of time [26, 36]. These subjects would contribute person-time only from that time forward. While the subcohort may be selected based on covariates, a key feature of the case–cohort design is that the subcohort is chosen without regard to failure status; methods that rely on failure status in the sampling of the comparison group are **case–control studies**.

Examples

Study of Lung Cancer Mortality in Aluminum Production Workers in Quebec, Canada. Armstrong et al. [1] describe the results of a case–cohort study selected from among 16 297 men who had worked at least one year in manual jobs at a large aluminum production plant between 1950 and 1988. This study greatly expands on an earlier cohort mortality study of the plant, which found a suggestion of increased rates of lung cancer in jobs with high exposures to coal tar pitch [12]. Through a variety of methods, 338 lung cancer deaths were identified. To avoid the expense associated with tracing subjects and abstraction of work records for the entire cohort, a case–cohort study was undertaken. To improve study efficiency a subcohort of 1138 subjects was randomly sampled from within year-of-birth strata with sampling fractions varying to yield a similar distribution to that of cases. This was accommodated in the analysis by stratification by these year-of-birth categories. The random sampling of subcohort members resulted in the inclusion of 205 cases in the subcohort. Work and smoking histories were abstracted for the subcohort and the additional 133 nonsubcohort cases. Cumulative exposure to coal tar pitch volatiles was estimated by linking worker job histories to measurements of chemical levels made in the plant using a **"job-exposure matrix"**. The analyses confirmed the lung cancer–coal pitch **association** observed in the earlier study and effectively ruled out confounding by smoking.

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

Women's Health Trial. To assess the potential health benefits of a low fat diet, a randomized trial of women assigned to low fat intervention and control groups has been undertaken. Of particular interest is the effect of this intervention on the risk of breast cancer. The study, as described in Self et al. [30], includes a cohort of 32 000 women between ages 45 and 69 whose percent calories from fat is greater than the **median** and who have at least one of a list of known risk factors for breast cancer. The study will involve 20 clinics across the US for a period of 10 years of follow-up. At two-year intervals, each participant will fill out four-day food records and food frequency questionnaires and blood will be drawn and stored. While evaluation of the intervention will be based on the full cohort, questions that require abstraction and coding of the questionnaires and blood lipid analyses are being addressed in a case–cohort study with a 10% sample serving as the subcohort. It was calculated that, relative to the entire cohort, this sample avoids about 80% of the cost of the analyses requiring these data with only a modest reduction of efficiency. The subcohort can also be used for making other comparisons between intervention and control groups. For example, the case–cohort sample could be used to investigate the joint relationship of blood hormone and nutrient levels and dietary intakes to breast cancer risk. Also, questions relating to other outcomes, such as cardiovascular disease, could be explored using the same subcohort as the comparison group, although additional data processing would be required for cases that occur outside the subcohort.

Statistical Analysis

Several methods have been developed to analyze case–cohort samples. Essentially, each of the methods available for the analysis of complete cohort data has an analog for the case–cohort sample. For point **estimation** of rate ratio parameters, the **likelihood** for full cohort data applied to the case–cohort data yields a valid estimator. However, estimation of the **variance** of point estimates, or tests of hypotheses (*see* **Hypothesis Testing**), requires adjustment to the standard full cohort variance estimators, as these will be too small. For likelihood-based methods, case–cohort sampling induces a covariance between score terms so that the variance of the **score** is given by $\Sigma + \Delta$, where Σ is the full cohort score variance and Δ is the sum of the covariances between the score terms. Since the subgroup used to compute the score terms has less variability than the full cohort, this covariance is positive. This leads to a larger variance for the parameter estimates, taking into account the subcohort sampling variability [19, 26, 29, 36]. Estimation of **absolute rates or risk** requires incorporation of the subcohort sampling fraction (or fractions) into the estimator.

Pseudo-likelihoods for Proportional Hazards Models

Assume the underlying model for disease rates has a **multiplicative** form:

$$\lambda[t, z(t); \beta_0] = \lambda_0(t) r[z(t); \beta_0],$$

where $r[z(t); \beta_0]$ is the rate 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 of disease in subjects with $z = 0$. The pseudo-likelihood approach described by Prentice [26] parallels the **partial likelihood** approach to the analysis of full cohort data. We start with the full cohort situation and then return to the analysis of the case–cohort sample. The partial likelihood approach is illustrated in Figure 1 for a small hypothetical **cohort study** of 15 subjects. Each horizontal line represents one subject. A 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**, i.e. be alive at the end of the study, died never having had the disease of interest, or be lost to follow-up. At each failure time a *risk set* is formed that includes the *case*; namely, the failure at that failure time, and all *controls*, namely, any other cohort members who are at risk at the failure time (these are denoted by a "|" in Figure 1). The partial likelihood for full cohort data is based on the **conditional probabilities** that the case failed given that one of the subjects in the risk set failed at that time. With r_k the rate ratio and Y_k the "at risk" indicator for subject k at the failure time, and r_{case} the rate ratio associated with the case, the full cohort partial likelihood is

$$\prod_{\text{failure times}} \frac{r_{\text{case}}}{\sum_{\text{case and all controls}} Y_k r_k}.$$

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

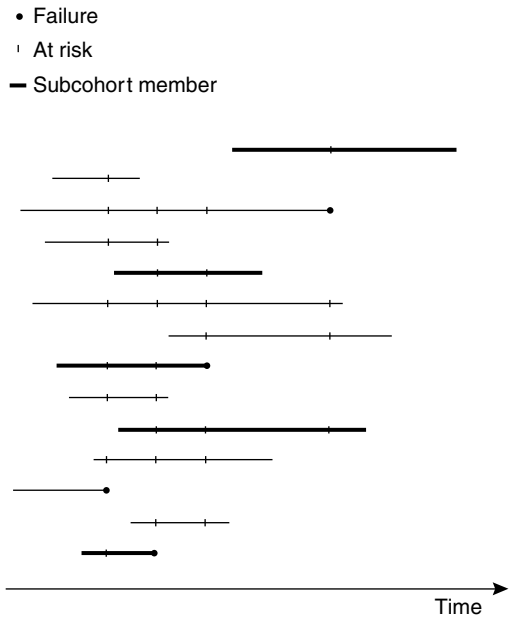

Figure 1 Prentice pseudo-likelihood approach to the analysis of case–cohort data. Pseudo-likelihood contributions are conditional probabilities based on the case and the subcohort members at risk at the failure time

Now in a case–cohort sample, covariate information is obtained for the subcohort and all nonsubcohort failures and only these subjects can contribute to the analysis. Prentice's pseudo-likelihood approach is illustrated in Figure 1 in which subcohort members are denoted by a thick horizontal line. For each failure, a *sampled risk set* is formed by the case and the controls who are in the subcohort (those with thick lines and a "|" at the failure time). As the figure indicates, subcohort members contribute to the analysis over their entire time on study, but the nonsubcohort failures contribute only at their failure times. Analogous to the full cohort partial likelihood, a pseudo-likelihood contribution is based on the conditional probability that the case fails given that someone fails among those in the sampled risk set. The pseudo-likelihood is then the product of such conditional probabilities over failure times:

$$\prod_{\text{failure times}} \frac{r_{\text{case}}}{\sum_{\text{case and subcohort controls}} Y_k r_k}, \qquad (1)$$

where the sum in the denominator is over the subcohort members when the case is in the subcohort and over the subcohort and nonsubcohort case when the case is not in the subcohort. This "likelihood" has the property that the expected value of the score is zero at β_0 but, as discussed above, the inverse information does not estimate the variance of the maximum pseudo-likelihood estimator. Prentice provided an estimator of the covariance Δ from the covariance between each pair of score terms, conditional on whether or not the failure occurring later in time was in the subcohort [17]. This is a rather complicated expression and only one software package has implemented it (Epicure, Hirosoft International Corp., Seattle, WA). Development of other methods of variance estimation has been an area of much research. These include "large sample" [29], **bootstrap** [37], "empirical" [9, 28], and influence function based [2, 21] methods. Simpler alternatives are the "asymptotic" [29] and the "robust" estimators [2, 21, 22]. Either may be computed by the simple manipulation of *delta beta* diagnostic statistics, which are an output option in many software packages [31]. The asymptotic estimator requires the sampling fractions, while the robust version estimates these from the data. Other methods of variance estimation have been proposed [9, 28, 37].

Absolute Risk Estimation

Estimation of the cumulative baseline hazard and related quantities parallel the nonparametric estimators based on the **Nelson–Aalen estimator** for full cohort data. Since the subcohort is a random sample from the full cohort, a natural estimator of the cumulative baseline hazard $\int_0^t \lambda_0(u)\,\mathrm{d}u$ is given by summing contributions for failure times up to t of the form:

$$\frac{1}{1/f \sum_{\text{subcohort}} r_k(\hat{\beta})},$$

where f is the proportion of the cohort in the subcohort [26, 29]. Again, adjustment of the cohort variance estimator is required. Cause-specific baseline hazard estimates for multiple outcomes have also been developed [25].

Other estimation methods and further developments

Alternative pseudo-likelihoods for the estimation of rate ratios of a similar form to (1) have been proposed. These involve differential weightings of the

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

r_k terms on the basis of the sampling fractions of those associated with subject k [2, 13, 29]. Iterative mean score methods have been proposed, which may yield more efficient estimators than (1) [8]. A method for analysis of generalized case–cohort sampling imputes rate ratio values for each cohort member using a "local averaging". Theoretically, this method is shown to have a superior efficiency to other methods, with methods for making the estimator optimal. Further research is needed to ascertain whether the increases are of practical importance. Methods for estimating standardized mortality ratios (*see* **Standardization Methods**) with a case–cohort sample have been described [35]. These involve "boosting up" the subcohort person-time in each age-year-exposure group "cell" by the inverse sampling fraction. Methods of variance adjustment are also discussed. When disease is rare and there is little censoring, methods of analyses for case-base studies with simple binary outcome data [23] will approximate the failure time analyses (e.g. [11, 17, 27, 33]). When exposure (or treatment) information is available on cohort members and additional information is to be collected for the case–cohort sample, an exposure-stratified subcohort may offer substantial efficiency advantages over random sampling. The analysis of this design uses a weighted variation of the pseudo-likelihood (1) and a generalization of the asymptotic variance estimator has been described [4].

Asymptotic Properties and Efficiency

Self & Prentice [29] give conditions for the **consistency** and asymptotic **normality** of the Prentice pseudo-likelihood for simple (stratified) case–cohort sampling. They show that the asymptotic variance of the maximum pseudo-likelihood estimator of relative risk parameters has the form $\Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}$, where Σ is the full cohort variance of the score, and they provide a formula for the asymptotic sampling-induced covariance Δ. This covariance depends on the censoring distribution even when $\beta_0 = 0$, so that efficiencies relative to the full cohort analysis must take the censoring distribution into account. Assuming a cohort with complete follow-up over a fixed observation period, an **exponential** relative risk model for a single binary covariate, a subcohort that is a simple $100\alpha\%$ random sample of the cohort, and probability of failure during the observation period of d, they calculate the **asymptotic relative efficiency** as

$$\left\{1 + 2\frac{1-\alpha}{\alpha}\left[1 + \frac{1-d}{d}\log(1-d)\right]\right\}^{-1}.$$

A number of papers have derived the asymptotic variance and semiparametric efficiency bounds for the case–cohort design [7, 8, 37, 38]. These indicate that, although the pseudo-likelihood (1) is not generally semiparametric efficient, the potential loss of efficiency appears to be small, unless disease is common or the size of the subcohort is much smaller than the number of cases.

Comparison with Nested Case–Control Sampling

Nested case–control and case–cohort methods are the two main approaches to sampling from assembled cohort studies. The former takes a retrospective point of view by sampling time-matched controls after the outcome (failure) occurs. In contrast, case–cohort sampling is prospective and unmatched in the sense that the comparison group, the subcohort, is picked without regard to failure status. Considerations for choosing between the designs have been the subject of some interest [10, 18, 20, 24, 26, 32, 34]. We summarize some of these considerations below.

Prospective Studies

If the study is **retrospective** and has been assembled, the major consideration in choosing between sampling designs is the statistical efficiency for the proposed analyses and the information to be collected on the sample, as this will translate quite directly into cost. If the study is prospective in that the study group will be assembled as time passes and outcomes occur in the future, the decision about which design to choose will depend on whether it is advantageous to have a comparison group early on in the study or whether it is better to wait until near the end. If the sample is to be chosen at the beginning, or concurrent with accrual into a prospective study, the case–cohort study has a number of advantages. First, as discussed above, processing of covariate information for the subcohort may proceed early on in the study during the accrual period. During the follow-up period, data for cases arising outside the

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

subcohort could be processed in batches at various times. A nested case–control study requires waiting until cases occur and controls are selected for them, delaying the processing of covariate information until later in the study than would be required in the case–cohort design. Thus, the case–cohort study can potentially be completed sooner than the nested case–control study. Secondly, although subcohort members should not be treated differently from cases occurring outside the subcohort, the subcohort can serve as a sample for assessing compliance, or quality control, as the study proceeds. However, a nested case–control sample may be advantageous if it is important that processing of information be "blinded" (*see* **Blinding or Masking**) to case-comparison group status. Since case–control covariate information can be processed simultaneously, potential information bias can be avoided. This is not always possible with a case–cohort sample when subcohort data are processed early in the study.

Statistical Efficiency

Comparison of statistical efficiency for studying a single outcome has been a topic of much research. It has been conjectured that the case–cohort design should be more efficient than the nested case–control design. This belief has been based on a comparison of the contribution of a failure to the pseudo-likelihood (1) with that of the corresponding nested case–control contribution. The former uses all subcohort members at risk at the failure time, whereas the latter uses only the controls selected for that case, usually resulting in the case–cohort having many more "controls per case". In fact, analytic and empirical efficiency comparisons indicate that in most situations encountered in practice, nested case–control sampling will be more efficient than the case–cohort, although often not by a large amount [18, 19, 34, 36]. The reason for the lower-than-anticipated relative efficiency is that the large number of controls per case in the case–cohort sample, which by itself increases efficiency, is offset by the sampling-induced positive **correlation** in score terms (see above), which lowers efficiency. The nested case–control design has relatively few controls per case, but there is no sampling-induced correlation between score terms [19].

Multiple Disease Outcomes

Since the subcohort is chosen without regard to failure status, it may serve as the comparison group for multiple disease outcomes. This would seem to be a great advantage over the nested case–control design, since controls are selected for specific cases. In fact, there are few published studies that exploit this feature of the design. Nevertheless, the most cost-effective use of the case–cohort design would seem to be to study a single set of explanatory factors and multiple outcomes. Thus, it seems likely that the case–cohort design may have application in clinical investigations in which researchers are often interested in multiple-event outcomes such as relapse, local and distant recurrence, and death as a function of a single set of treatment and **prognostic factors**. If, for instance, the prognostic factors involve expensive laboratory work, a case–cohort sample would be a natural way to reduce costs associated with the laboratory work, but still allow a full analysis of multiple endpoints. Using the same comparison group will result in correlation between estimates of the same parameter for different endpoints. Appropriate methods for the variance adjustment and hypothesis testing with multiple outcomes have been developed [25].

Matching

Often, it is desirable to match (*see* **Matching**) comparison subjects closely on certain factors, either to control for **confounding** or so that information of comparable quality may be obtained. For instance, it is common to compare a case with controls close in year of birth to adjust for secular trends in behavior. Fine matching, and matching based on time-dependent factors is accommodated in a natural way in a nested case–control sample. Matching may only be done crudely for case–cohort sampling and must be based on factors available at the time the subcohort is sampled.

Analysis Flexibility

The nested case–control design is inherently associated with methods for analysis of cohort data based on **semiparametric proportional hazards** models. Estimation of rate ratio parameters is based on partial likelihood methods and estimation of absolute-risk-related quantities is based on the Nelson–Aalen

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

estimator of the **cumulative hazard**. (One interesting exception to the restriction to proportional hazards models is estimation of excess risks using the **Aalen linear model** [3].) The case–cohort design is not associated with any particular model or method of analysis. Thus, in theory, "Poisson likelihood" or "grouped time" case–base analysis approaches, as well as the risk-set-based pseudo-likelihood (1), may be used for parameter estimation. Examples of estimation of parameters in nonproportional hazards models from case–cohort data include the additive hazards, proportional odds, and transformation regression models [5, 6, 14]. For a subcohort that is a simple random sample, changing time scales and analysis stratification variables poses no difficulties in the analysis. Since the nested case–control sample is bound to the risk set defined by the time scale and stratification variables used in matching controls to failures, these must be fixed in the analysis. However, **inference** from case–cohort samples is complicated by the need to adjust **standard errors** and test statistics for the sampling-induced covariance. Further, for testing, adjusted Wald and score tests are adapted in a natural way using the variance estimator [31], but a "pseudo-likelihood ratio test" is not available.

Computation

Standard conditional logistic regression software, for the analysis of matched case–control data, may be used to analyze rate ratio parameters from nested case–control studies (*see* **Software, Biostatistical**). Furthermore, if the numbers of subjects in the risk sets are known, absolute risk estimators and standard errors are relatively simple to compute [16]. Since the latter are based on standard nonparametric cumulative hazard and survival estimators, standard software for the analysis of full cohort data may be "tricked" into computing the nested case–control estimators. For case–cohort samples, standard **Cox regression** software may be used to estimate parameters but, as discussed above, special software is needed to estimate corresponding variances.

References

[1] Armstrong, B., Tremblay, C., Baris, D. & Gilles, T. (1994). Lung cancer mortality and polynuclear aromatic hydrocarbons: a case–cohort study of aluminum production workers in Arvida, Quebec, Canada, *American Journal of Epidemiology* **139**, 250–262.

[2] Barlow, W.E. (1994). Robust variance estimation for the case–cohort design, *Biometrics* **50**, 1064–1072.

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

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

[5] Chen, H.Y. 2001. Fitting semiparametric transformation regression models to data from a modified case-cohort design, *Biometrika* **88**(1), 255–268.

[6] Chen, H.Y. 2001. Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design, *Journal of the American Statistical Association* **96**(456), 1446–1457.

[7] Chen, Kani 2001. Generalized case-cohort sampling, *Journal of the Royal Statistical Society, Series B, Methodological* **63**(4), 791–809.

[8] Chen, K. & Lo, S-H. 1999. Case-cohort and case-control analysis with Cox's model, *Biometrika* **86**(4), 755–764.

[9] Edwardes, M.D. (1995). Re: Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality, *Statistics in Medicine* **14**, 1609–1610.

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

[11] Flanders, W.D., Dersimonian, R. & Rhodes, P. (1990). Estimation of risk ratios in case–base studies with competing risks, *Statistics in Medicine* **9**, 423–435.

[12] Gibbs, G.W. (1985). Mortality of aluminum reduction plant workers, 1950 through 1977, *Journal of Occupational Medicine* **27**, 761–770.

[13] Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality, *Statistics in Medicine* **7**, 149–160.

[14] Kulich, M. & Lin,D.Y. 2000. Additive hazards regression for case-cohort studies, *Biometrika* **87**, 73–87.

[15] 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.

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

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

[18] Langholz, B. & Thomas, D.C. (1990). Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison, *American Journal of Epidemiology* **131**, 169–176.

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

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022

[20] Langholz, B., Thomas, D.C., Witte, J.S. & Peters, R.K. (1995). Re: Thompson et al. a population based case–cohort evaluation of the efficacy of mammography screening for breast cancer, *American Journal of Epidemiology* **142**, 448–449.

[21] Lin, D.Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**, 1341–1349.

[22] Mark, S.D. & Katki, H. 2001. Influence function based variance estimation and missing data issues in case-cohort studies, *Lifetime Data Analysis* **7**(4), 331–344.

[23] Miettinen, O.S. (1982). Design options in epidemiology research: an update, *Scandinavian Journal of Work, Environment, and Health* **8**(Supplement 1), 1295–1311.

[24] Moulton, L.H., Wolff, M.C., Brenneman, G. & Santosham, M. (1995). Case–cohort analysis of case-coverage studies of vaccine effectiveness, *American Journal of Epidemiology* **142**, 1000–1006.

[25] Sorensen Per & Andersen, Per Kragh 2000. Competing risks analysis of the case-cohort design, *Biometrika* **87**(1), 49–59.

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

[27] Sato, T. (1992). Maximum likelihood estimation of the risk ratio in case–cohort studies, *Biometrics* **48**, 1215–1221.

[28] Schouten, E.G., Dekker, J.M., Kok, F.J., Le Cessie, S., van Houwelingen, H.C., Pool, J. & Vandenbrouke, J.P. (1993). Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality, *Statistics in Medicine* **12**, 1733–1745.

[29] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case–cohort studies, *Annals of Statistics* **16**, 64–81.

[30] Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S.L., Clifford, C., Goldman, S., Urban, N., Sheppard, L. & Greenwald, P. (1988). Statistical design of the women's health trial, *Controlled Clinical Trials* **9**, 119–136.

[31] Therneau, T.M. & Li, H. 1999. Computing the Cox model for case cohort designs, *Lifetime Data Analysis* **5**, 99–112.

[32] Thompson, R.S., Barlow, W.E., Taplin, S.H., Grothaus, L., Immanuel, V., Salazar, A. & Wagner, E.H. (1994). A population-based case–cohort evaluation of the efficacy of mammographic screening for breast cancer, *American Journal of Epidemiology* **140**, 889–901.

[33] van den Brandt, P.A., Goldbohm, R.A. & van't Veer, P. (1995). Alcohol and breast cancer: results from the Netherlands cohort study, *American Journal of Epidemiology* **141**, 907–915.

[34] Wacholder, S. (1991). Practical considerations in choosing between the case–cohort and nested case–control designs, *Epidemiology* **2**, 155–158.

[35] Wacholder, S. & Boivin, J.-F. (1987). External comparisons with the case–cohort design, *American Journal of Epidemiology* **126**, 1198–1209.

[36] 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.

[37] Wacholder, S., Gail, M.H., Pee, D. & Brookmeyer, R. (1989). Alternative variance and efficiency calculations for the case–cohort design, *Biometrika* **76**, 117–123.

[38] Zhang, H. & Goldstein, L. 2002. Information and asymptotic efficiency of the case-cohort sampling design in Cox's regression model, *Journal of Multivariate Analysis* in press.

BRYAN LANGHOLZ

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 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.b2a03022