Two contributions to the update of the Encyclopedia of Statistical Sciences: Nested case-control sampling and Counter-matched sampling

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

May, 1997

Abstract

This report contains two contributions to the update of the Encyclopedia of Statistical Sciences. The contributions in this report are

NESTED CASE-CONTROL SAMPLING (SAMPLING FROM THE RISK SETS)

Cox's regression model* is one of the cornerstones in modern survival analysis*, and it is the method of choice when one wants to assess the influence of risk factors and other covariates on mortality or morbidity. Estimation in Cox's model is based on a partial likelihood* which, at each observed death or disease occurrence ("failure"), compares the covariate values of the failing individual to those of all individuals at risk at the time of the failure. In large epidemiological cohort studies of a rare disease (see EPIDEMIOLOGICAL STATISTICS and COHORT ANALYSIS), Cox regression requires collection of covariate information on all individuals in the cohort even though only a small fraction of these actually get diseased. This may be very expensive, or even logistically impossible. Cohort sampling techniques where covariate information is collected for all failing individuals ("cases"), but only for a sample of the non-failing individuals ("controls") then offer useful alternatives which may drastically reduce the resources that need to be allocated to a study. Further, as most of the statistical information is contained in the cases, such studies may still be sufficient to give reliable answers to the questions of interest.

The most common cohort sampling design is nested case-control sampling. Here one compares each case to a small number of controls selected at random from those at risk at the case's failure time, and a new sample of controls is selected for each case. A different type of cohort sampling design is case-cohort sampling. For this design one selects at the outset of the study a random sample of control individuals (the subcohort), and these individuals are used as controls throughout the study (provided they are still at risk). In this entry we focus on the nested case-control design. We first indicate the relation between this form of case-control sampling and the more classical case-control designs (see RETROSPECTIVE STUDIES, INCLUDING CASE-CONTROL) and give a sketch of the development of the subject. To fix ideas, we then describe in more details one particular nested case-control study. Further, we review the Cox model, describe precisely how the nested case-control data are collected, and present methods for statistical inference. Finally a note on efficiency is given, and we provide some remarks on extensions of the nested case-control design as well as a brief comparison between nested case-control sampling and case-cohort sampling.

Nested case-control studies and other case-control designs

The theory for case-control studies for a binary response variable (diseased, not diseased) dates back to the work of Cornfield [9] in the early 1950s (see ODDS RATIO ESTIMATORS), proceeds via the landmark 1959 paper by Mantel and Haenzel [14](see MANTEL-HAENSZEL STATISTIC) to the implementation of the logistic regression* model and the development of conditional logistic regression for matched case-control data in the 1970s. The monograph by Breslow and Day [7] gives an extensive exposition of this "classical" case-control theory, while [6] provides a nice historical account.

Age or other time-scales play no role in the statistical models on which the "classical" case-control theory is based, so this important aspect of a study has to be taken care of by stratification or time-matching. This is different for a nested case-control study, where

Cox's regression model is used to model the occurrence of failures, and where the controls are sampled from the risk sets. Further, in a "classical" case-control study the population from which the controls are sampled is often not well defined, while a nested case-control study is performed within a well defined cohort. This makes the nested case-control design intermediate between a "classical" case-control study and a full cohort analysis.

The nested case-control design was suggested in 1977 by Thomas [18] as a tool to reduce error checking and the computational burden for the analysis of large cohorts. He proposed to base inference on a modification of Cox's partial likelihood, see (6) below. This suggestion was supported by the work of Prentice and Breslow [17], who derived the same expression as a conditional likelihood for time-matched case-control sampling from an infinite population. A more decisive, but still heuristic, argument was provided by Oakes [15], who showed that (6) is a partial likelihood when the sampling of controls is performed within the actual finite cohort. It took more than ten years, however, before Goldstein and Langholz [10] proved rigorously that the estimator of the regression coefficients based on Oakes' partial likelihood enjoys similar large sample properties as ordinary maximum likelihood estimators. Later Borgan, Goldstein and Langholz [3] gave a more direct proof along the lines of Andersen and Gill [2] using a marked-point process formulation. It is indicated below how this marked point process approach also solves the problem of how to estimate the baseline hazard rate function from nested case-control data.

An example

The nested case-control design has been used in many studies to avoid the collection of covariate information for the full cohort or to reduce error checking and the computational burden in the analysis of large cohorts. In fact, it is now recognized that most time-matched case-controls studies, ubiquitous in epidemiological research, are indeed nested case-control studies where the cohort is given as the (sometimes not well defined) population within a given geographic area. In order to fix ideas in the subsequent discussion, we will have a closer look at one such study.

The International Agency for Research on Cancer in Lyon, France maintains a register of 21 183 workers from eleven countries (Australia, Austria, Canada, Denmark, Finland, Germany, Italy, the Netherlands, New Zealand, Sweden, and the United Kingdom) exposed to phenoxy herbicides, chlorophenols, and dioxins. In a cohort analysis, an increased mortality of soft tissue sarcomas was found among exposed subjects. In order to examine the effect of exposure to various chemicals more fully, a nested case-control study was undertaken, where, for each of the 11 cases of soft tissue sarcoma (all males), five controls were sampled at random from those from the same country and of the same age as the case [11]. The degree of the exposures of the 11 cases and the 55 controls to a number of chemicals were reconstructed through the use of individual job records and of detailed company exposure questionnaires and company reports. Even though this information in principle could have be collected for the complete cohort of 21 183 individuals, this would have implied an enormous amount of work. Further, as the main limitation of the data is the small number of cases, such an effort would mostly have been in vain. In fact, the nested case-control study based on the 11+55 controls and cases provides 83% efficiency relative to the full cohort data for

testing associations between single exposures and the disease (cf. below).

Model and data

Consider a cohort of n individuals and denote by $\lambda_i(t) = \lambda(t; \mathbf{x}_i(t))$ the hazard rate function at time t for an individual i with vector of covariates $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$. Here the time-variable t may be age (as in the example), time since employment, or some other time-scale relevant to the problem at hand. The covariates may be time-fixed (like gender) or time-dependent (like cumulative exposure to a chemical), and they may be indicators for categoric covariates (like the exposure groups "non-exposed," "low," "medium," and "high") or numeric (as when actual amount of exposure is recorded). Cox's regression* model relates the covariates of individual i to its hazard rate function by

$$\lambda_i(t) = \lambda_0(t) e^{\boldsymbol{\beta}' \mathbf{x}_i(t)}. \tag{1}$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients, while the baseline hazard rate function $\lambda_0(t)$, corresponding to an individual with all covariates identically equal to zero, is left unspecified.

Sometimes one adopts a stratified version of (1) where the baseline hazard rate function may differ between strata, while the regression coefficients are assumed the same across strata. E.g., for the situation discussed in the example, it may be reasonable to stratify according to country of residence since both exposure and occurrence or recognition of disease may differ by country. In order to simplify the presentation, we will concentrate on the model (1) with no stratification, and only comment upon the modifications for the situation with stratification when relevant.

The individuals in the cohort may be followed over different periods of time, i.e., our observations may be subject to left-truncation and/or right censoring (see TRUNCATION and CENSORED DATA). The risk set $\mathcal{R}(t)$ is the collection of all individuals who are under observation just before time t, and $n(t) = \#\mathcal{R}(t)$ is the number at risk at that time. We let $t_1 < t_2 < \cdots$ be the times when failures are observed and, assuming that there are no tied failures, denote by i_j the index of the individual who fails at t_j (a few ties may be broken at random). A nested case-control sample is then obtained as follows: At each failure time t_j, one selects by simple random sampling* without replacement $m-1$ individuals (controls) from the $n(t_j) - 1$ non-failing individuals in $\mathcal{R}(t_j)$. The sampled risk $\widetilde{\mathcal{R}}(t_j)$ then consists of the case i_j and these $m-1$ controls. Covariate information is collected for all individuals in the sampled risk sets, but are not needed for the remaining individuals in the cohort. Note that the sampling is done independently at the different failure times, so an individual may be member of more than one sampled risk set.

A basic assumption is that truncation and censoring, as well as the sampling of controls, are independent in the sense that the additional knowledge of which individuals have entered the study, have been censored or have been selected as controls before any time t do not carry information on the risks of failure at t; cf. Sections III.2-3 in [1] and [3] for a general discussion. For a small time-interval $[t, t + \mathrm{d}t)$, this assumption and (1) imply that

$$\Pr\left(i \text{ fails in } [t, t+\mathrm{d}t) \mid \mathcal{F}_{t-}\right) = e^{\boldsymbol{\beta}' \mathbf{x}_i(t)} \lambda_0(t) \mathrm{d}t \tag{2}$$

if individual i is at risk just before time t. Here "the history" $\mathcal{F}_{t-}$ contains information about observed failures, entries, exits and changes in covariate values in the cohort, as well as information on the sampling of controls, up to but not including time t. (Not all this information will actually be available to the researcher in a nested-case control study.)

In the example, controls were not selected from all individuals at risk, but only from those at risk in the same country as the case. (Since age is used as time-scale when forming the risk sets, the controls will also have the same age as the case.) This way of sampling the controls is related to the stratified Cox model discussed earlier. When the stratified model applies, the sampling of controls should be restricted to those at risk in the same stratum as the case. We say that the controls are matched by the stratification variable (country in the example).

Estimation

Estimation of the regression coefficients in (1) is based on a partial likelihood* which may be derived heuristically as follows. Denote by $\widetilde{\mathcal{R}}(t)$ the sampled risk set were a failure to occur at t, and let $\mathcal{P}(t)$ be the collection of all possible sampled risk sets at that time. Then $\mathcal{P}(t)$ is the set of all $\binom{n(t)}{m}$ subsets of $\mathcal{R}(t)$ of size m. Consider a set $\mathbf{r} \in \mathcal{P}(t)$ and an individual $i \in \mathbf{r}$. Then by (2), and since the $m-1$ controls are sampled by simple random sampling without replacement from the $n(t)-1$ non-failing individuals in $\mathcal{R}(t)$,

$$\begin{aligned}
&\Pr\left(i \text{ fails in } [t, t+\mathrm{d}t),\, \widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \mathcal{F}_{t-}\right) \\
&\quad = \Pr\left(\widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, i \text{ fails at } t,\, \mathcal{F}_{t-}\right) \times \Pr\left(i \text{ fails in } [t, t+\mathrm{d}t) \,|\, \mathcal{F}_{t-}\right) \\
&\quad = \binom{n(t)-1}{m-1}^{-1} e^{\boldsymbol{\beta}' \mathbf{x}_i(t)} \lambda_0(t) \mathrm{d}t. \qquad (3)
\end{aligned}$$

Now the sampled risk set equals $\mathbf{r}$ if one of the m individuals in $\mathbf{r}$ fails, and the remaining $m-1$ individuals are selected as controls. Therefore

$$\Pr\left(\text{one individual in } \mathbf{r} \text{ fails in } [t, t+\mathrm{d}t),\, \widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \mathcal{F}_{t-}\right) = \binom{n(t)-1}{m-1}^{-1} \sum_{l \in \mathbf{r}} e^{\boldsymbol{\beta}' \mathbf{x}_l(t)} \lambda_0(t) \mathrm{d}t. \qquad (4)$$

Dividing (3) by (4), it follows that

$$\Pr\left(i \text{ fails at } t \;\middle|\; \text{one individual in } \mathbf{r} \text{ fails at } t,\, \widetilde{\mathcal{R}}(t) = \mathbf{r},\, \mathcal{F}_{t-}\right) = \frac{e^{\boldsymbol{\beta}' \mathbf{x}_i(t)}}{\sum_{l \in \mathbf{r}} e^{\boldsymbol{\beta}' \mathbf{x}_l(t)}}. \qquad (5)$$

Multiplying together conditional probabilities of the form (5) for all failure times t_j, cases i_j, and sampled risk sets $\widetilde{\mathcal{R}}(t_j)$, we arrive at the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{t_j} \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{i_j}(t_j)}}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} e^{\boldsymbol{\beta}' \mathbf{x}_l(t_j)}}. \qquad (6)$$

This is similar to the full cohort partial likelihood (see COX'S REGRESSION MODEL) except that the sum in the denominator is taken over the sampled risk set $\widetilde{\mathcal{R}}(t_j)$ in place of

the full cohort risk set $\mathcal{R}(t_j)$. Inference concerning $\boldsymbol{\beta}$, using the usual large-sample likelihood methods, can be based on the partial likelihood (6). The above heuristic derivation of (6) is essentially the one given by Oakes [15] . Borgan, Goldstein and Langholz [3] made this argument rigorous using a marked point processes formulation.

The partial likelihood (6) also applies for the stratified Cox model when the controls are matched by the stratification variable, i.e., sampled from the same stratum as the case. Note further that (6) is of the same form as the conditional likelihood for logistic regression with $m-1$ matched controls per case (e.g., Chapter 7 in [7]). Thus standard software for conditional logistic regression can be used for data analysis.

The cumulative baseline hazard rate function $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$ can be estimated by [3, 4]

$$\widehat{\Lambda}_0(t) = \sum_{t_j \leq t} \frac{1}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} e^{\widehat{\boldsymbol{\beta}}' \mathbf{x}_l(t_j)} \frac{n(t_j)}{m}}, \tag{7}$$

where $\widehat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimator maximizing (6). The estimator (7) is of the same form as the one used for cohort data (see SURVIVAL ANALYSIS). However, since nested case-control data only use information from a sample of those at risk, the contribution for each subject in the sampled risk set, including the case, is weighted by the inverse of the proportion sampled. The estimator (7) is almost unbiased when averaged over all possible failure and sampled risk set occurrences.

When there is only a small number of strata, the stratum specific cumulative baseline hazard rate functions for the stratified Cox model may be estimated by a slight modification of (7) . All that is required is that the sum is restricted to those failure times t_j when a failure in the actual stratum occurs, and that the $n(t_j)$ are taken to be the number at risk in this stratum. When there are many strata, however, there may be too little information in each stratum to make estimation of the stratum specific cumulative baseline hazard rate functions meaningful.

Relative efficiency

Goldstein and Langholz [10] were the first to carry out a rigorous study of the asymptotic properties of the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ maximizing (6). Based on the asymptotic distributional results, they also presented a study of the asymptotic efficiency of the maximum partial likelihood estimator for nested case-control data relative to the estimator based on the full cohort partial likelihood. When $\boldsymbol{\beta} = \mathbf{0}$, the asymptotic covariance matrix of the nested case-control estimator equals $m/(m-1)$ times the asymptotic covariance matrix of the full cohort estimator, independent of censoring and covariate distributions. Thus the efficiency of the nested case-control design relative to the full cohort is $(m-1)/m$ for testing associations between single exposures and disease – a result which has been known for some time for binary covariates based on the time-matched case-control study paradigm [8]. In the example with 5 controls per case this yields the relative efficiency of 5/6=83% mentioned earlier.

When $\boldsymbol{\beta}$ departs from zero, and when more than one regression coefficient has to be estimated, the efficiency of the nested case-control design may be much lower than given by

the "$(m-1)/m$ efficiency rule" [8, 10]. E.g., with one binary covariate for exposure with relative risk $e^{\beta} = 4$, the relative efficiency of the nested case-control design with one control per case is about 1/4 when 10% of the cohort is exposed rather than 1/2 as the rule suggests.

Properly normalized the estimator (7) for the cumulative baseline hazard rate function converges weakly to a Gaussian process [3]. This asymptotic distributional result makes it possible to study the asymptotic efficiency of (7) relative to the full cohort estimator, but such an efficiency-study has yet to be performed. Preliminary studies by the author of this entry indicate, however, that the relative efficiency of (7) is much higher than the one of the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$.

Extensions

Estimation of the cumulative hazard rate function for an individual with given time-fixed covariate values were studied in [4] and extended to time-varying covariate histories in [13]. The latter also described how the results presented earlier extend to regression models where (1) is replaced by $\lambda_i(t) = \lambda_0(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t))$ for some relative risk function $r(\boldsymbol{\beta}, \mathbf{x}_i(t))$ and discussed estimation of absolute risk without and in the presence of competing risks*. Estimation of excess risk from nested case-control data using Aalen's nonparametric linear regression model was discussed in [5].

In a nested case control study, the controls are selected by simple random sampling. An alternative is to select the controls by stratified random sampling (see STRATIFIED DESIGNS). This design, termed counter-matched sampling*, may reduce the estimation uncertainty in situations of practical interest [12]. Using a marked-point process formulation, a general framework for the sampling of controls incorporating the nested case control design and counter-matched sampling as special cases, were introduced and studied in [3].

A study design related to nested case-control studies, is the case-cohort design [16]. Here a subcohort $\mathcal{C}$ is selected by simple random sampling from the entire cohort at the outset of the study. Covariate information is collected for the members of $\mathcal{C}$ as well as for cases occurring outside this subcohort. Estimation of $\boldsymbol{\beta}$ is based on a pseudo likelihood which has the same form as (6), but with $\widetilde{\mathcal{R}}(t_j)$ replaced by $(\mathcal{C} \cap \mathcal{R}(t_j)) \cup \{i_j\}$ for the sums in the denominator. Since this estimation is not based on a partial likelihood, the usual large sample likelihood methods do not apply, and this makes the analysis of data from a case-cohort study more cumbersome than the analysis of nested case-control data. For most studies involving a single disease, the case-cohort and nested case-control design seem to have about the same efficiency. The main potential of the case-cohort design therefore lies in situations where multiple disease endpoints are to be evaluated, because disease free members of the subcohort may serve as controls for the disease cases of each type.

References

[1] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

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

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

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

[5] Borgan, Ø. and Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model. *Biometrics* **53**, 10-17.

[6] Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14–28.

[7] Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research.* Volume 1 – *The Analysis of Case-Control Studies*, *IARC Scientific Publications*, Vol. 32. International Agency for Research on Cancer, Lyon.

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

[9] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11**, 1269–1275.

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

[11] 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., 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.

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

[13] Langholz, B. and Borgan, Ø. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53**, 44-51.

[14] Mantel, N. and Haenzel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

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

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

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

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

(BIOSTATISTICS
COUNTER-MATCHED SAMPLING
COX'S REGRESSION MODEL
EPIDEMIOLOGICAL STATISTICS
PARTIAL LIKELIHOOD
RETROSPECTIVE STUDIES (INCLUDING CASE-CONTROL)
SURVIVAL ANALYSIS)

COUNTER-MATCHED SAMPLING

Counter-matching is a novel design for statified sampling of controls in epidemiological case-control studies. It is a generalization of nested case-control sampling* and will often give an efficiency gain compared to this classical design.

Counter-matched sampling and the nested case-control design are closely related to Cox's regression model* for failure time data. This model relates the vector of covariates $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ at time t for an individual i to its hazard rate function $\lambda_i(t)$ by

$$\lambda_i(t) = \lambda_0(t) e^{\boldsymbol{\beta}' \mathbf{x}_i(t)}. \tag{8}$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients, while the baseline hazard rate function $\lambda_0(t)$ is left unspecified. Estimation in Cox's model is based on a partial likelihood* which, at each failure time, compares the covariate values of the failing individual to those of all individuals at risk at the time of the failure. In large epidemiological cohort studies of a rare disease (see EPIDEMIOLOGICAL STATISTICS and COHORT ANALYSIS), Cox regression requires the collection of information on exposure variables and other covariates of interest for all individuals in the cohort even though only a small fraction of these actually get diseased. This may be very expensive, or even logistically impossible. Nested case-control studies, in which covariate information is needed only for each failing individual ("case") and a small number of controls selected from those at risk at the time of the failure, may give a substantial reduction in the resources required for a study. Moreover, as most of the statistical information is contained in the cases, a nested case-control study may still be sufficient to give reliable answers to the questions of main interest.

In the classical form of a case-control study nested within a cohort, the controls are selected by simple random sampling* (see NESTED CASE-CONTROL SAMPLING). Often some information is available for all cohort members, e.g., a surrogate measure of exposure, like type of work or duration of employment, may be available for everyone. Langholz and Borgan [3] have developed a stratified version of the simple nested case-control design which makes it possible to incorporate such information into the sampling process in order to obtain a more informative sample of controls. For this design, called *counter-matching*, one applies the additional information on the cohort subjects to classify each individual at risk into one of say, L, strata. Then at each failure time t_j, one samples randomly without replacement m_l controls from the $n_l(t_j)$ at risk in stratum l, except for the case's stratum where only $m_l - 1$ controls are sampled. The failing individual i_j is, however, included in the sampled risk set $\widetilde{\mathcal{R}}(t_j)$, so this contains a total of m_l from each stratum $l = 1, 2, \ldots, L$. In particular, for $L = 2$ and $m_1 = m_2 = 1$ the single control is selected from the opposite stratum of the case. Thus counter-matching is, as the name suggests, essentially the opposite of matching where the case and its controls are from the same stratum.

Inference from counter-matched data concerning $\boldsymbol{\beta}$ in (8) can be based on the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{t_j} \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{i_j}(t_j)} w_{i_j}(t_j)}{\sum_{k \in \widetilde{\mathcal{R}}(t_j)} e^{\boldsymbol{\beta}' \mathbf{x}_k(t_j)} w_k(t_j)} \tag{9}$$

using the usual large-sample likelihood methods [1, 3]. Here $w_k(t_j) = n_l(t_j)/m_l$ if individual k belongs to stratum l at time t_j. The partial likelihood (9) is similar to Oakes' [6] partial

likelihood for simple nested case-control data (see NESTED CASE-CONTROL SAMPLING). But the contribution of each individual, including the case, has to be weighted by the inverse of the proportion sampled from the individual's stratum in order to compensate for the different sampling probabilities in the strata. The cumulative baseline hazard rate function $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$ can be estimated by [1]

$$\widehat{\Lambda}_0(t) = \sum_{t_j \le t} \frac{1}{\sum_{k \in \widetilde{\mathcal{R}}(t_j)} e^{\widehat{\boldsymbol{\beta}}' \mathbf{x}_k(t_j)} w_k(t_j)}, \tag{10}$$

where $\widehat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimator maximizing (9). The estimator (10) is also similar to the one used for nested case-control data.

Counter-matching may give an appreciable improvement in statistical efficiency for estimation of a regression coefficient of particular importance compared to simple nested case-control sampling. Intuitively this is achieved by increasing the variation in the covariate of interest within each sampled risk set. The efficiency gain has been documented both by asymptotic relative efficiency calculations [3, 4, 5] and by Steenland and Deedens' [7] study of a cohort of gold miners. For the latter, a counter-matched design (with stratification based on duration of exposure) with three controls per case had the same statistical efficiency for estimating the effect of exposure to crystalline silica as a simple nested case-control study using ten controls. According to preliminary investigations by the author of this entry, a similar increase in efficiency is not seen for the estimator (10). One important reason for this is that for estimation of the baseline hazard rate function even a nested case-control study has quite high efficiency compared to the full cohort.

The idea of counter-matching originated in the middle of the nineties of the 20th century and is rather new at the time of this writing (1997). It has therefore not yet been put into practical use. But it has attracted positive interest from researchers in epidemiology [2, 7], and it is quite likely to be a useful design for future epidemiological studies.

References

[1] Borgan, Ø., Goldstein L., and Langholz (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*, **23**, 1749–1778. (The paper uses marked point processes to describe a general framework for risk set sampling designs, including simple nested case-control sampling and counter-matched sampling as special cases. Large sample properties of the estimators of the regression coefficients and the cumulative baseline hazard rate function are studied using counting process and martingale theory.)

[2] Cologne, J. B. (1997). Counter-intuitive matching. *Epidemiology*, in press. (Invited editorial advocating the use of counter-matching for an epidemiological audience.)

[3] Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82**, 69-79. (The basic paper where the concept of counter-matching was introduced and studied. Comparisons with simple nested case-control sampling are also provided.)

[4] Langholz, B. and Clayton, D. (1994). Sampling strategies in nested case-control studies. *Environmental Health Perspectives* **102 (Suppl 8)**, 47-51. (A non-technical paper which discusses a number of practical situations where counter-matching may be useful. Comparisons with simple nested case-control sampling are also provided.)

[5] Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science*, **11**, 35-53. (The paper reviews a broad variety of risk set sampling designs, including simple nested case-control sampling and counter-matched sampling, and discusses when these are appropriate for different design and analysis problems from epidemiologic research.)

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

[7] Steenland, K. and Deedens, J. A. (1997). Estimating exposure-response trends in nested case-control studies: control selection via counter-matching versus random sampling. *Epidemiology*, in press. (An applied paper which compares counter-matching to simple nested case-control sampling for a real data set.)

(BIOSTATISTICS
COX'S REGRESSION MODEL
EPIDEMIOLOGICAL STATISTICS
NESTED CASE-CONTROL SAMPLING
PARTIAL LIKELIHOOD
SURVIVAL ANALYSIS)