Methods for the analysis of sampled cohort data in the Cox proportional hazards model

Ø. Borgan
University of Oslo

L. Goldstein
University of Southern California

B. Langholz
University of Southern California

Abstract

Methods are provided for regression parameter and cumulative baseline hazard estimation in the Cox model when the cohort is sampled according to a predictable sampling probability law. It is shown how a marked point process representation of cohort sampling naturally leads to the derivation of a partial likelihood which may be used for the estimation of regression parameters. Standard counting process techniques are used to show that this partial likelihood may be treated as a likelihood in that, at the true parameter, the expectation of the score is zero and the variance of the score is the expected information. Generalizations of the Breslow estimator of the cumulative baseline hazard and estimators of the associated variance processes are provided. The results are used to derive partial likelihoods for three new sampling designs, stratified and quota sampling from the risk sets and nested case-control sampling with number of controls dependent on the failure's exposure status, as well as for simple nested case-control and case-cohort sampling. Baseline hazard estimators are given for simple and stratified nested case-control sampling. General asymptotic theory is developed for the maximum partial likelihood estimator and cumulative baseline hazard estimator and is used to derive the asymptotic distributions for estimators from simple and stratified nested case-control sampling. Generalisations to stratified populations or multistate problems and the Aalen linear regression model are given.

1 Introduction

Epidemiologic cohort studies are considered the most reliable method for assessing the variation in rates of morbidity and mortality due to factors present in the population under study. Cohort members are observed over some time period, and either "fail" (develop or die from the disease of interest) or are "censored" (are alive at the end of the study period, die of some other cause, or are lost to follow-up). Variation in rates are then modeled from information on exposures, confounders, and other potential predictors of risk, which are generically called "covariates," collected on cohort members. If complete covariate information is obtained for all cohort members, a wide range of parametric and semi-parametric analytic techniques are available (e.g. Breslow and Day, 1987). Especially

[0] *AMS 1980 subject classifications:* 62D05, 62F12, 62G05, 62M99, 62P10

[0] *Key words:* Aalen's linear regression model; Case-control study; Cohort sampling; Counting process; Cox's regression model; Epidemiology; Marked point process; Martingale; Partial likelihood; Survival analysis

[0] *Abbreviated title:* Cohort sampling methods

useful has been the semi-parametric proportional hazards model (Cox, 1972), where the hazard rate for the ith subject is specified as

$$\alpha_i(t) = \alpha_0(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t)). \tag{1.1}$$

Here $\mathbf{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip}(t))^{\mathsf{T}}$ is a vector of covariates for subject i at time t, β_0 is a vector of regression parameters, and α_0 is the baseline hazard about which minimal assumptions are made. The partial likelihood used for estimation of the regression parameters does not depend α_0 and is given as

$$\mathcal{L}(\beta) = \prod_{t_j}\left\{\frac{\exp(\beta^{\mathsf{T}}\mathbf{Z}_{i_j}(t_j))}{\sum_{l\in\mathcal{R}_j}\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(t_j))}\right\}, \tag{1.2}$$

where the t_j are the ordered failure times, i_j is the index of the failure at time t_j and $\mathcal{R}_j$ is the set of all those "at risk" at t_j, the failure and those on study. An estimator of the cumulative baseline hazard, $A_0(t) = \int_0^t \alpha_0(t)\mathrm{d}t$, was given by Breslow (1972) as

$$\widehat{A}_0(t;\widehat{\beta}) = \sum_{t_j\le t}\frac{1}{\sum_{l\in\mathcal{R}_j}\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{Z}_l(t_j))} \tag{1.3}$$

with $\widehat{\beta}$ the maximum partial likelihood estimator obtained by maximizing (1.2). Andersen and Gill (1982) showed that (1.2) has "basic likelihood properties," i.e., that the expectation of the score at the true value β_0 of the regression parameters is zero and the variance of the score at β_0 is the inverse information, and developed the asymptotic theory for the estimators $\widehat{\beta}$ and $\widehat{A}_0$. These estimators provide a basis for inference about other quantities or functions of interest such as survival probabilities and median survival times (e.g. Andersen *et al.*, 1983; Andersen *et al.*, 1992, Section VII.2.3), and slight modifications of these provide estimators for e.g. relative mortality (Andersen *et al.*, 1985) or general relative risk parameters (Prentice and Self, 1983). Further, though we have, and will continue to use terminology which indicates that a subject may only fail once, these methods apply, without modification, to outcomes which may recur and have been extended to stratified populations and multistate models (Andersen and Borgan, 1985, Section 7).

Typically, because of the rarity of the disease outcome and/or the complexity of the relationships to be explored, cohort studies require very large numbers of subjects and/or long periods of follow up in order to accumulate enough failures to have sufficient statistical power to give reliable answers to the questions of interest. This leads to something of a paradox. If a cohort study is large enough to allow for a meaningful analysis, the cost of collecting high quality covariate information on all subjects is prohibitively expensive, if not logistically impossible. It would also seem unnecessary. Loosely, if the disease of interest is rare, the contribution of the non-failures, in terms of the "power" of the study, will be negligible compared to that of the failures. Thus cohort sampling methods which include all the failures and a portion of the non-failures are highly desirable.

Perhaps surprisingly, there exist few cohort sampling options. The most popular of these is the nested case-control design (Thomas, 1977) and this is of particular interest in the context of this paper because, with the exception of a never used sampling-with-replacement variant (Robins *et al.*, 1986), it is the only cohort sampling method which is analyzed using partial likelihood techniques. Thus, at this point, we will focus our attention on nested case-control sampling and return to other sampling methods suggested in

the literature in Section 10. In the nested case-control design, sampled risk sets $\tilde{\mathcal{R}}_j$ consist of the "case" (failure) at t_j and $m-1$ "controls" randomly selected from those at risk at t_j (and, possibly meet some other matching criteria). Considering that time (on some scale) matched population based case-control studies generally are nested case-control samples from a large cohort usually loosely defined by geographic region, this design is indeed ubiquitous in epidemiologic research. The nested case-control partial likelihood has the same form as that for the full cohort except that $\tilde{\mathcal{R}}_j$ replaces $\mathcal{R}_j$ in (1.2) (Oakes, 1981) and, under suitable conditions, may be treated like an ordinary likelihood (Goldstein and Langholz, 1992). The efficiency of this design relative to the full cohort for testing for an association between a single covariate and disease is $(m-1)/m$. Thus, there is little to be gained by sampling more than 4 or 5 controls if detecting simple associations is the goal of the study. However, it is becoming increasingly rare that the aims of epidemiologic studies are so modest. Often of interest are the manner in which rates change with increasing exposure, assessing the effect of exposure after adjustment for confounders, and the variation of the effect of exposure as a function of potential "effect modifying" factors. In such situations, indications are that the efficiency of nested case-control sampling can be much lower than for the simple test of association (Breslow *et al.*, 1983). As the questions considered grow in complexity and the costs involved increase, it may prove very advantageous to tailor the sampling design to take into account the goals of the study and/or "recognize" the costly aspects of the study. In principle, the design should attain a specified level of (relative) efficiency for the analysis goal(s) of interest at the lowest "cost." This principle is purely academic if statistical methods to analyze data collected using such designs do not exist. While it is beyond the scope of this paper to go into these issues in detail, we note that such considerations have resulted in development of new sampling designs for unmatched case-control studies including stratified sampling (Fears and Brown, 1986; Scott and Wild, 1991), two-stage sampling (Breslow and Cain, 1988), and randomized recruitment (Weinberg and Wacholder, 1990).

In this paper, we develop methods for the analysis of a large class of cohort sampling designs which parallel those available for full cohort data. Estimation of the regression parameters for a given design is surprisingly simple; it is based on maximizing a partial likelihood, defined precisely in Section 3, which is proportional to

$$\mathcal{L}(\beta) = \prod_{t_j} \left\{ \frac{\exp(\beta^{\mathsf{T}} \mathbf{Z}_{i_j}(t_j)) w_{i_j}(t_j)}{\sum_{l \in \tilde{\mathcal{R}}_j} \exp(\beta^{\mathsf{T}} \mathbf{Z}_l(t_j)) w_l(t_j)} \right\}, \tag{1.4}$$

where $\tilde{\mathcal{R}}_j$ is the sampled risk set and the $w_l(t)$ are (simple) weight functions which depend on the sampling method. (Henceforth, we will not make any distinction between the partial likelihood and functions which are proportional to it.) Under conditions given in Section 3, (1.4) is shown to have basic likelihood properties and, in Section 6, additional conditions are given to ensure the consistency and asymptotic normality of the maximum partial likelihood estimator $\hat{\beta}$ obtained by maximizing (1.4). Thus, analysis of sampled data is particularly simple, standard conditional logistic regression software, used for the analysis of matched case-control studies, accommodates (1.4) without modification. The only additional requirement is that an "offset," a term in the linear predictor for which no regression parameter is estimated, of $\log w_l(t_j)$ must be added to the model. This feature is currently available in most packages designed for the analysis of epidemiologic studies.

Thusfar, case-cohort sampling is the only sampling method for which a cumulative baseline hazard estimator has been provided (Self and Prentice, 1988). In Section 4, we

give a natural extension of the full cohort baseline hazard estimator (1.3) for sampled cohort data. In particular, in Section 5.2, we provide the estimator for nested case-control sampling. This provides a basis for relatively straightforward sampled data estimators of various quantities and functions which parallel those available for the full cohort.

The key to the development of these methods is to use a marked point process to model simultaneously events happening in the cohort (like failures and censorings) and the sampling of controls at each failure time. Heuristically speaking, a marked point process $\{(t_j, x_j); j \geq 1\}$ is just an ordered sequence of time point $t_1 < t_2 < \cdots$ where events occur together with marks $x_1, x_2, \ldots$ which describe the events happening at these times. The marks x_j take values in a set E called the *mark space*. In the general theory of marked point processes (e.g. Brémaud, 1981; Karr, 1986) the mark space may be very general. We will, however, only consider marked point processes with a finite mark space. For the class of sampling designs we shall consider, a mark of the form $x_j = (i_j, \tilde{\mathcal{R}}_j)$ will indicate that the individual with index i_j is failing at t_j and that $\tilde{\mathcal{R}}_j$ is the sampled risk set at that time. Here $\tilde{\mathcal{R}}_j$ consists of the case i_j together with its sampled set of controls.

The outline of the paper is as follows. Section 2 gives the general marked point process model for sampling of risk sets and the class of sampling schemes covered by this methodology. Counting and intensity processes, used in the partial likelihood construction, are presented along with the associated martingale structure. In Section 3, we derive a partial likelihood for cohort sampling and show that it has basic likelihood properties and, in Section 4 derive an estimator for the underlying integrated baseline hazard for a restricted class of sampling schemes. Along with the full cohort data and simple nested case-control sampling, applications given in Section 5 include three new, and potentially quite useful, sampling designs: stratified and quota sampling extensions of the simple nested case-control design and nested case-control sampling with variable matching ratio. Case-cohort sampling also belongs to this class and is given as an example where the maximum partial likelihood estimator is clearly inefficient. Formal derivations of the large sample properties of maximum partial likelihood and cumulative hazard estimators are given in Sections 6 and, in Section 7, we apply these results to some of the applications presented in Section 5. Sections 8 and 9 are devoted to extending our methodology to generalizations of the simple proportional hazards model (1.1). Specifically, Section 8 develops methods for multistate models and stratified populations and Section 9 for the Aalen linear regression model. In the final Section 10, we briefly discuss efficiency issues and the relationship of this paper to earlier work on cohort sampling methods.

Throughout the paper we will, without further reference, use standard results from the theory of multivariate counting processes, local square integrable martingales and stochastic integrals as surveyed e.g. by Fleming and Harrington (1991, Chapters 1-2) and Andersen *et al.* (1992, Sections II.2-4). We will only consider marked point processes with a finite mark space, so we do not, however, need results on marked point processes beyond those surveyed by Arjas (1989, Sections 2 and 4) and Andersen *et al.* (1992, Sections II.4 and II.7).

2 A marked point process model for sampled cohort data

We fix throughout the paper a time interval $[0, \tau]$ for a given terminal time τ, $0 < \tau \leq \infty$.

First we specify a model for events observed in the cohort, without consideration of the sampling of controls, along the lines of Andersen *et al.* (1992, Section III.5). Let the cohort

consist of n individuals, and assume that all events observed to happen, that is registered failures as well as information on individuals entering or leaving the study population and on observed changes in time dependent covariate values, may be modeled as a marked point process $\{(t_j^*, x_j^*); j \geq 1\}$ on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The filtration generated by this marked point process, together with covariate values at $t = 0$, is denoted $(\mathcal{H}_t)_{t\in[0,\tau]}$. This is an increasing, right-continuous family of sub-σ-algebras of $\mathcal{F}$, and $\mathcal{H}_t$ specifies the "cohort history" up to time t in the sense that it contains all events (up to null sets) whose occurrence or not is fixed by time t. There is also a pre-t σ-algebra $\mathcal{H}_{t-}$ which contains all events whose occurrence or not is fixed strictly before time t.

From the marked point process $\{(t_j^*, x_j^*); j \geq 1\}$ we may extract a marked point process $\{(t_j, i_j); j \geq 1\}$ which only records the innovative marks, i.e. the times t_j when failures occur and the individuals i_j which fail at these time points. Of course the marked point process $\{(t_j, i_j); j \geq 1\}$ is adapted to $(\mathcal{H}_t)$. Associated with this marked point process we have the counting processes

$$N_i(t) = \sum_{j\geq 1} I(t_j \leq t, i_j = i) \tag{2.1}$$

counting the number of observed failures for individual i in $[0, t]$, $i = 1, 2, \ldots, n$. We also introduce $Y_i(t)$ for the predictable indicator process taking the value 1 if the ith individual is at risk at $t-$ and 0 otherwise.

A model for the cohort is now given by relating the intensity process λ_i for N_i to the vector $\mathbf{Z}_i(t)$ of (possibly) time dependent covariates for the ith individual; $i = 1, 2, \ldots, n$. (The model is only partially specified since we do not specify models for the censoring mechanism and the covariate processes.) A fundamental assumption we make is that the $\mathbf{Z}_i(t)$ are left continuous and adapted; consequently they are predictable and locally bounded. In particular, this means that the values of the covariates at time t should be known, based on available information on the cohort, just before time t. We will consider a proportional hazards regression model, where the intensity processes are specified as

$$\lambda_i(t) = Y_i(t)\alpha_0(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t)), \tag{2.2}$$

with $\alpha_0(t)$ a non-negative baseline intensity or hazard function and β_0 a p-dimensional vector (Cox, 1972; Andersen and Gill, 1982). In fact, at the cost of somewhat more complicated proofs in Section 6, our results may be extended to relative risk regression models where $\exp(\cdot)$ in (2.2) is replaced by a positive relative risk function $r(\cdot)$ standardized so that $r(0) = 1$.

Now that a model for the cohort has been specified, we turn to describe how the sampling of controls is superimposed onto this model. This is done by sampling at each failure time t_j (according to a distribution to be specified below) a set of controls for the failing individual i_j. We denote by $\tilde{\mathcal{R}}_j$ the sampled risk set consisting of the set of these controls together with the individual i_j failing at t_j. Then

$$\{(t_j, (i_j, \tilde{\mathcal{R}}_j)); j \geq 1\} \tag{2.3}$$

will be a marked point process with a finite mark space E which may be specified as follows: Let $\mathcal{P}$ be the power set of $\{1, 2, \ldots, n\}$, i.e. the set of all subsets of $\{1, 2, \ldots, n\}$, and let

$$\mathcal{P}_i = \{\mathbf{r} : \mathbf{r} \in \mathcal{P}, i \in \mathbf{r}\}.$$

Then the mark space of (2.3) is given by

$$E = \{(i, \mathbf{r}) : i \in \{1, 2, \ldots, n\}, \mathbf{r} \in \mathcal{P}_i\} = \{(i, \mathbf{r}) : \mathbf{r} \in \mathcal{P}, i \in \mathbf{r}\}.$$

The introduction of the sampling into the model will bring in some extra random variation, so the marked point process (2.3) will not be adapted to the filtration $(\mathcal{H}_t)$ generated by the available data from the cohort. Thus we now have to work with the enlarged family of sub-σ-algebras $(\mathcal{F}_t)_{t \in [0,\tau]}$ of $\mathcal{F}$ given by

$$\mathcal{F}_t = \mathcal{H}_t \vee \sigma\{\tilde{\mathcal{R}}_j;\, t_j \leq t\},$$

i.e. $(\mathcal{F}_t)$ is generated by the observed events in the cohort together with the sampled risk sets.

Corresponding to the marked point process (2.3) we now have, for each $(i, \mathbf{r}) \in E$, the counting process

$$N_{(i,\mathbf{r})}(t) = \sum_{j \geq 1} I(t_j \leq t, (i_j, \mathcal{R}_j) = (i, \mathbf{r})) \tag{2.4}$$

counting the observed number of failures for the ith individual in $[0, t]$ with associated sampled risk set $\mathbf{r}$. Since the mark space E is finite, the marked point process (2.3) is, in fact, equivalent to the multivariate counting process $(N_{(i,\mathbf{r})};\, (i, \mathbf{r}) \in E)$. We denote the intensity process of $N_{(i,\mathbf{r})}$ by $\lambda_{(i,\mathbf{r})}$. From (2.4) we may recover the counting process (2.1), registering the observed failures for the ith individual, by summing over all possible sampled risk sets, i.e.

$$N_i(t) = \sum_{\mathbf{r} \in \mathcal{P}_i} N_{(i,\mathbf{r})}(t), \tag{2.5}$$

and a similar relation holds for the intensity processes λ_i and $\lambda_{(i,\mathbf{r})}$.

The fact that we now have to consider the filtration $(\mathcal{F}_t)$, also containing information about the sampled risk sets, may have the consequence that the intensity processes corresponding to the counting processes N_i may change, i.e. their $(\mathcal{F}_t)$-intensity processes may differ from their $(\mathcal{H}_t)$-intensity processes (2.2). For instance, in a prevention trial, this will be the case if individuals selected as controls change their behavior in such a way that their risk of failure is different from similar individuals which have not been previously selected as controls. To rule out such possibilities we need the concept of *independent sampling* analogous to the usual assumption that censoring must be independent (Andersen *et al.*, 1992, Section III.2.2).

Formally, we will say that we have *independent sampling provided that the $(\mathcal{F}_t)$-intensity processes of the counting processes N_i are the same as their $(\mathcal{H}_t)$-intensity processes.* In other words: the additional knowledge of sampling which has occurred before any time t should not alter the intensities of failure at t. We note that under independent sampling the $(\mathcal{F}_t)$-intensity processes of the N_i are given by (2.2). In the following we will tacitly assume that the sampling is independent. Further, we will consider intensity processes, martingales, etc. with respect to the filtration $(\mathcal{F}_t)$, and not the "cohort history" $(\mathcal{H}_t)$.

Then, given $\pi_t(\mathbf{r} | i)$, the conditional probability of selecting the sampled risk set $\mathbf{r} \in \mathcal{P}_i$ at time t given $\mathcal{F}_{t-}$ and the fact that the ith individual fails at t, a model for the marked

point process (2.3) may be given by specifying the intensity processes $\lambda_{(i,\mathbf{r})}$ for the counting processes (2.4) by

$$\lambda_{(i,\mathbf{r})}(t) = \lambda_i(t)\pi_t(\mathbf{r}|\, i). \tag{2.6}$$

Note that the $\pi_t(\mathbf{r}|i)$ may be recovered by

$$\pi_t(\mathbf{r}|\, i) = \frac{\lambda_{(i,\mathbf{r})}(t)}{\lambda_i(t)} = \frac{\lambda_{(i,\mathbf{r})}(t)}{\sum_{\mathbf{r}\in\mathcal{P}_i}\lambda_{(i,\mathbf{r})}(t)}.$$

For notational convenience we set $\pi_t(\mathbf{r}|i) = 0$ if $Y_i(t) = 0$.

Thus by (2.2)

$$\lambda_{(i,\mathbf{r})}(t) = Y_i(t)\alpha_0(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t))\pi_t(\mathbf{r}|\, i), \tag{2.7}$$

and it follows that a model for cohort sampling is given by specifying, for each t and each i with $Y_i(t) = 1$, the sampling distributions $\pi_t(\cdot|\, i)$ over sets $\mathbf{r}$ in $\mathcal{P}_i$. This specification must be based on information available just before time t, i.e. the $\pi_t(\mathbf{r}|\, i)$ must be predictable considered as processes in t (for fixed i and $\mathbf{r}$). In particular this rules out selection of controls depending on events in the future, e.g. one may not exclude as potential controls for a current case individuals that *subsequently* fail (Lubin and Gail, 1984). We will give examples of specific sampling distributions in Section 5.

By standard counting process theory it follows that for $(i,\mathbf{r}) \in E$,

$$M_{(i,\mathbf{r})}(t) = N_{(i,\mathbf{r})}(t) - \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \tag{2.8}$$

are local square integrable martingales. Their predictable variation processes are given as

$$\langle M_{(i,\mathbf{r})}\rangle(t) = \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u, \tag{2.9}$$

while their predictable covariation processes are

$$\langle M_{(i,\mathbf{r})}, M_{(j,\mathbf{s})}\rangle(t) = 0 \tag{2.10}$$

for $(i,\mathbf{r}) \neq (j,\mathbf{s})$.

In the derivation of a partial likelihood in the next section, we will need to consider the reduced marked point process

$$\{(t_j, \tilde{\mathcal{R}}_j);\, j \geq 1\} \tag{2.11}$$

derived from (2.3) by disregarding the information about which individuals fail at the various time points. Corresponding to this marked point process we have the counting processes

$$N_{\mathbf{r}}(t) = \sum_{i\in\mathbf{r}} N_{(i,\mathbf{r})}(t) \tag{2.12}$$

counting the number of times the sampled risk set equals $\mathbf{r}$ in $[0,t]$. By (2.7) these have intensity processes

$$\lambda_{\mathbf{r}}(t) = \sum_{i\in\mathbf{r}} \lambda_{(i,\mathbf{r})}(t) = \sum_{i\in\mathbf{r}} Y_i(t)\alpha_0(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t))\pi_t(\mathbf{r}|\, i). \tag{2.13}$$

Moreover,

$$M_{\mathbf{r}}(t) = N_{\mathbf{r}}(t) - \int_0^t \lambda_{\mathbf{r}}(u)\mathrm{d}u, \tag{2.14}$$

for $\mathbf{r} \in \mathcal{P}$, are local square integrable martingales with

$$\langle M_{\mathbf{r}} \rangle(t) = \int_0^t \lambda_{\mathbf{r}}(u)\mathrm{d}u, \tag{2.15}$$

and

$$\langle M_{\mathbf{r}}, M_{\mathbf{s}} \rangle(t) = 0 \tag{2.16}$$

for $\mathbf{r} \neq \mathbf{s}$.

We conclude this section with two remarks concerning our model construction. First, for ease of presentation, we have assumed that censoring is a part of the cohort history $(\mathcal{H}_t)$ only, and that no extra censoring is introduced by the nested case-control sampling. This may easily be extended, however, along the lines of Andersen *et al.* (1992; Section III.2) to include extra (independent) censoring depending on the previous sampling history. For example, in a nested case-control study, one may censor individuals after they have been picked as controls.

Second, in order to define our model, we had to augment the cohort history $(\mathcal{H}_t)$ with the sampling history in order to get the filtration $(\mathcal{F}_t)$ relative to which intensity processes, martingales, etc. are defined. Thus more "information" is needed to define the model for nested case-control sampling than is the case for the full cohort. At first this may seem somewhat paradoxical since the main reason for sampling is that one does not need to collect covariate information for all individuals in the cohort. But the paradox is resolved when one remembers that the model must describe the likelihood of *all* possible outcomes, not just the outcome actually observed. The fact that more "information" is needed to define the model than what is actually observed, has the consequence, however, that care must be exercised in order to make sure that statistical inference procedures are based only on data which are actually available to the researcher.

3 A partial likelihood and estimation of the regression parameter

In the previous section we derived a (partially specified) probabilistic model for nested case-control sampling. We now consider the statistical model obtained by allowing α_0 and β_0 to be varying parameters. We will assume that the baseline hazard α_0 may be an arbitrary non-negative function while the vector of regression parameters β_0 takes values in the p-dimensional Eucledian space.

As described in general terms in Andersen *et al.* (1992; Section II.7.3) the full likelihood may be factorised into the (partial) likelihood for the marked point process (2.3) and a second factor depending on the (unspecified) model for the censoring mechanism and the covariate processes. We use α and β for the free parameters in likelihoods, etc. and reserve α_0 and β_0 for the true values of these parameters. Then the likelihood for (2.3) takes the form

$$\prod_{u\in[0,\tau]} \prod_{\mathbf{r}\in\mathcal{P}} \prod_{i\in\mathbf{r}} \left(\lambda_{(i,\mathbf{r})}(u;\alpha,\beta)^{\Delta N_{(i,\mathbf{r})}(u)}\right) \exp\left(-\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \sum_{i\in\mathbf{r}} \lambda_{(i,\mathbf{r})}(u;\alpha,\beta)\mathrm{d}u\right), \tag{3.1}$$

where we have written $\lambda_{(i,\mathbf{r})}(t;\alpha,\beta)$ for the intensity processes (2.7) in order to emphasize their dependence on the parameters.

Now inference cannot be based directly on (3.1) for two reasons. First, our model is semi-parametric in the sense that α may be any non-negative function. So, as for the classical Cox model, (3.1) may be made arbitrarily large by letting α be zero except very close to the failure times t_j, where we let it peak higher and higher. Second, in order to be able to evaluate the second term of (3.1) we would need to know the full covariate histories for all individuals in the cohort. But as discussed at the end of Section 2, we do not actually collect all this covariate information when the cohort is sampled.

We therefore consider a partial likelihood (Cox, 1975) which may be obtained using the general ideas of Arjas (1989; Section 4). To this end we factorise the intensity processes $\lambda_{(i,\mathbf{r})}$, not as in (2.6), but as

$$\lambda_{(i,\mathbf{r})}(t) = \lambda_{\mathbf{r}}(t)\pi_t(i\,|\,\mathbf{r};\beta_0),$$

where, by (2.7) and (2.13),

$$\pi_t(i\,|\,\mathbf{r};\beta_0) = \frac{\lambda_{(i,\mathbf{r})}(t)}{\lambda_{\mathbf{r}}(t)} = \frac{Y_i(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t))\pi_t(\mathbf{r}|\,i)}{\sum_{l\in\mathbf{r}} Y_l(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_l(t))\pi_t(\mathbf{r}|\,l)} \tag{3.2}$$

is the conditional probability of the ith individual failing at t, given $\mathcal{F}_{t-}$ and that there is a failure among individuals in the set $\mathbf{r}$ at t. Note that (3.2) does not depend on the baseline hazard.

Statistical inference on β may therefore be based on the partial likelihood

$$\begin{aligned}\mathcal{L}_\tau(\beta) &= \prod_{t_j}\pi_{t_j}(i_j\,|\,\tilde{\mathcal{R}}_j;\beta) = \prod_{t_j}\left\{\frac{Y_{i_j}(t_j)\exp(\beta^{\mathsf{T}}\mathbf{Z}_{i_j}(t_j))\pi_{t_j}(\tilde{\mathcal{R}}_j|\,i_j)}{\sum_{l\in\tilde{\mathcal{R}}_j} Y_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(t_j))\pi_{t_j}(\tilde{\mathcal{R}}_j|\,l)}\right\} \\ &= \prod_{u\in[0,\tau]}\prod_{\mathbf{r}\in\mathcal{P}}\prod_{i\in\mathbf{r}}\left\{\frac{Y_i(u)\exp(\beta^{\mathsf{T}}\mathbf{Z}_i(u))\pi_u(\mathbf{r}|\,i)}{\sum_{l\in\mathbf{r}} Y_l(u)\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(u))\pi_u(\mathbf{r}|\,l)}\right\}^{\Delta N_{(i,\mathbf{r})}(u)},\end{aligned} \tag{3.3}$$

obtained by only using the information contained in the conditional distributions of the failing individuals i_j given the sampled risk sets $\tilde{\mathcal{R}}_j$, and thereby disregarding the information on β contained in the reduced marked point process (2.11). This generalizes the partial likelihood of Oakes (1981) for nested case-control designs with simple random sampling of the controls (Example 5.2). Note that in the denominator of (3.3) each individual is weighted with the probability of selecting the sampled risk set had the individual been the failure. In particular, no distinction is made between the failure and the controls in the sampled risk set.

Note that by (2.12), (2.13) and (3.2) the likelihood (3.1) for the marked point process (2.3) may be factorised as

$$\prod_{u\in[0,\tau]}\prod_{\mathbf{r}\in\mathcal{P}}\prod_{i\in\mathbf{r}}\left((\lambda_{\mathbf{r}}(u;\alpha,\beta)\pi_u(i\,|\,\mathbf{r};\beta))^{\Delta N_{(i,\mathbf{r})}(u)}\right)\exp\left(-\int_0^\tau\sum_{\mathbf{r}\in\mathcal{P}}\lambda_{\mathbf{r}}(u;\alpha,\beta)\mathrm{d}u\right)$$

$$= \prod_{u\in[0,\tau]}\prod_{\mathbf{r}\in\mathcal{P}}\left(\lambda_{\mathbf{r}}(u;\alpha,\beta)^{\Delta N_{\mathbf{r}}(u)}\right)\exp\left(-\int_0^\tau\sum_{\mathbf{r}\in\mathcal{P}}\lambda_{\mathbf{r}}(u;\alpha,\beta)\mathrm{d}u\right)$$

$$\times\prod_{u\in[0,\tau]}\prod_{\mathbf{r}\in\mathcal{P}}\prod_{i\in\mathbf{r}}\pi_u(i\,|\,\mathbf{r};\beta)^{\Delta N_{(i,\mathbf{r})}(u)},$$

i.e. as a product of the likelihood for the reduced marked point process (2.11) and the partial likelihood (3.3).

The estimator $\widehat{\beta}$ obtained by maximizing the partial likelihood (3.3) has similar properties as a maximum likelihood estimator. At this point we will only show that (3.3) has basic likelihood properties, i.e. that the score vector has expectation zero and that its covariance matrix equals the expected information matrix (implicitly assuming the necessary regularity conditions to hold) and return to a detailed study in Section 6.

We introduce the notation

$$S_{\mathbf{r}}^{(\gamma)}(\beta,t)=\frac{\partial^\gamma}{\partial\beta^\gamma}S_{\mathbf{r}}^{(0)}(\beta,t)=\sum_{i\in\mathbf{r}}Y_i(t)\mathbf{Z}_i(t)^{\otimes\gamma}\exp(\beta^{\mathsf{T}}\mathbf{Z}_i(t))\pi_t(\mathbf{r}\,|\,i) \tag{3.4}$$

for $\gamma=0,1,2$, where for a vector $\mathbf{a}$, $\mathbf{a}^{\otimes 0}=1$, $\mathbf{a}^{\otimes 1}=\mathbf{a}$ and $\mathbf{a}^{\otimes 2}=\mathbf{a}\mathbf{a}^{\mathsf{T}}$. Note that $S_{\mathbf{r}}^{(1)}(\beta,t)$ is a p-vector and $S_{\mathbf{r}}^{(2)}(\beta,t)$ a $p\times p$-matrix. Furthermore, we define the p-vector

$$\mathbf{E}_{\mathbf{r}}(\beta,t)=S_{\mathbf{r}}^{(1)}(\beta,t)/S_{\mathbf{r}}^{(0)}(\beta,t) \tag{3.5}$$

and the $p\times p$-matrix

$$\mathbf{V}_{\mathbf{r}}(\beta,t)=\frac{S_{\mathbf{r}}^{(2)}(\beta,t)}{S_{\mathbf{r}}^{(0)}(\beta,t)}-\mathbf{E}_{\mathbf{r}}(\beta,t)^{\otimes 2}. \tag{3.6}$$

The two quantities $\mathbf{E}_{\mathbf{r}}(\beta,t)$ and $\mathbf{V}_{\mathbf{r}}(\beta,t)$ are the expectation and the covariance matrix, respectively, of the covariate vector $\mathbf{Z}_i(t)$ if an individual is selected with probability $\pi_t(i\,|\,\mathbf{r};\beta)$; cf. (3.2).

Apart from a constant term the log partial likelihood equals

$$C_\tau(\beta)=\int_0^\tau\sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}}\left\{\beta^{\mathsf{T}}\mathbf{Z}_i(u)-\log(S_{\mathbf{r}}^{(0)}(\beta,u))\right\}\mathrm{d}N_{(i,\mathbf{r})}(u), \tag{3.7}$$

and differentiation with respect to β yields the vector of score functions

$$\mathbf{U}_\tau(\beta)=\frac{\partial}{\partial\beta}C_\tau(\beta)=\int_0^\tau\sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}}\{\mathbf{Z}_i(u)-\mathbf{E}_{\mathbf{r}}(\beta,u)\}\,\mathrm{d}N_{(i,\mathbf{r})}(u) \tag{3.8}$$

and the observed information matrix

$$\mathcal{I}_\tau(\beta)=-\frac{\partial^2}{\partial\beta^2}C_\tau(\beta)=\int_0^\tau\sum_{\mathbf{r}\in\mathcal{P}}\mathbf{V}_{\mathbf{r}}(\beta,u)\mathrm{d}N_{\mathbf{r}}(u). \tag{3.9}$$

By the interpretation of $\mathbf{V}_{\mathbf{r}}(\beta,t)$ as a covariance matrix given just below (3.6), it follows that $\mathcal{I}_\tau(\beta)$ is positive definite, and hence that the log partial likelihood (3.7) is concave.

Using (2.7), (2.8), (3.4) and (3.5) it is seen that the score, evaluated at the true parameter value β_0, equals the (vector valued) stochastic integral

$$\mathbf{U}_\tau(\beta_0) = \int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \sum_{i\in\mathbf{r}} \{\mathbf{Z}_i(u) - \mathbf{E}_\mathbf{r}(\beta_0, u)\}\, \mathrm{d}M_{(i,\mathbf{r})}(u). \tag{3.10}$$

In particular it follows that the expected score is zero.

We let $\mathbf{U}_t(\beta_0)$ be defined by (3.10), but with the integral taken over $[0,t]$ instead of $[0,\tau]$, and note that $\mathbf{U}_\cdot(\beta_0)$ is a (vector valued) local square integrable martingale. By a standard argument, the matrix of predictable covariation processes of this martingale, evaluated at τ, becomes (use (2.7), (2.8)-(2.10) and (3.4)-(3.6))

$$\begin{aligned}\langle \mathbf{U}_\cdot(\beta_0)\rangle(\tau) &= \int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \sum_{i\in\mathbf{r}} \{\mathbf{Z}_i(u) - \mathbf{E}_\mathbf{r}(\beta_0, u)\}^{\otimes 2}\, \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \\ &= \int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \mathbf{V}_\mathbf{r}(\beta_0, u) S_\mathbf{r}^{(0)}(\beta_0, u)\alpha_0(u)\mathrm{d}u. \end{aligned} \tag{3.11}$$

Moreover, by (2.12), (2.14) and (3.4), the observed information matrix (3.9) evaluated at β_0 may be written as

$$\mathcal{I}_\tau(\beta_0) = \int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \mathbf{V}_\mathbf{r}(\beta_0, u) S_\mathbf{r}^{(0)}(\beta_0, u)\alpha_0(u)\mathrm{d}u + \int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} \mathbf{V}_\mathbf{r}(\beta_0, u)\mathrm{d}M_\mathbf{r}(u). \tag{3.12}$$

Thus the observed information matrix equals the predictable variation of the score plus a local square integrable martingale. In particular, by taking expectations, it follows that the expected information matrix equals the covariance matrix of the score.

4 Estimation of the integrated baseline hazard and survival probabilities

For completely general sampling distributions $\pi_t(\cdot\,|\,\mathbf{r})$, as considered in the previous section, it seems difficult to derive a sensible estimator for the integrated baseline hazard

$$A_0(t) = \int_0^t \alpha_0(t)\mathrm{d}t.$$

We will therefore now restrict ourselves to sampling distributions with a special structure which can be described as follows: Conditional on $\mathcal{F}_{t-}$ there exists for each t, at which there is at least one individual at risk, a sampling distribution $\pi_t(\cdot)$ over sets $\mathbf{r}$ in $\mathcal{P}$ such that

$$\pi_t(\mathbf{r}|\, i) = \frac{\pi_t(\mathbf{r})}{p_t(i)} \tag{4.1}$$

for $i \in \mathbf{r}$. Here

$$p_t(i) = \sum_{\mathbf{r}\in\mathcal{P}_i} \pi_t(\mathbf{r}) \tag{4.2}$$

is the probability that the individual i will be included in a sample selected according to $\pi_t(\cdot)$, and this probability is assumed to be positive for each individual at risk at t. Thus (4.1) specifies $\pi_t(\mathbf{r}|\, i)$ as the conditional probability of selecting the sample $\mathbf{r}$ at t given $\mathcal{F}_{t-}$ and that i is contained in the sample. In Section 5 we will illustrate how some important sampling schemes for the controls have this structure.

For sampling distributions satisfying (4.1), we suggest the estimator

$$\begin{aligned}\widehat{A}_0(t;\widehat{\beta}) &= \sum_{t_j \le t} \frac{1}{\sum_{l\in\widetilde{\mathcal{R}}_j} Y_l(t_j)\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{Z}_l(t_j))/p_{t_j}(l)} \\ &= \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \frac{\mathrm{d}N_{\mathbf{r}}(u)}{\sum_{l\in\mathbf{r}} Y_l(u)\exp(\widehat{\beta}^{\mathsf{T}}\mathbf{Z}_l(u))/p_u(l)},\end{aligned} \tag{4.3}$$

for the integrated baseline hazard $A_0(t)$. To motivate this estimator, let

$$J(t) = I\left(\sum_{i=1}^n Y_i(t) > 0\right) \tag{4.4}$$

be the predictable indicator process which equals 1 if someone is at risk at $t-$ and equals 0 otherwise, and remember that $\pi_t(\mathbf{r})$ is only defined when $J(t) = 1$. We interpret $\pi_t(\mathbf{r})J(t)$ as 0 when $J(t) = 0$. Then use (2.13), (2.14) and (4.1) to find that (4.3), with $\widehat{\beta}$ replaced by the true value β_0, may be written as

$$\begin{aligned}\widehat{A}_0(t;\beta_0) &= \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \frac{J(u)\mathrm{d}N_{\mathbf{r}}(u)}{\sum_{l\in\mathbf{r}} Y_l(u)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_l(u))/p_u(l)} \\ &= \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J(u)\alpha_0(u)\pi_u(\mathbf{r})\mathrm{d}u + \widehat{W}(t) \\ &= \int_0^t J(u)\alpha_0(u)\mathrm{d}u + \widehat{W}(t),\end{aligned} \tag{4.5}$$

where

$$\widehat{W}(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \frac{J(u)\mathrm{d}M_{\mathbf{r}}(u)}{\sum_{l\in\mathbf{r}} Y_l(u)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_l(u))/p_u(l)} \tag{4.6}$$

is a local square integrable martingale. This shows that $\widehat{A}_0(t;\beta_0)$ is almost unbiased for $A_0(t)$ (the small bias only being due to the possibility of having no one at risk), thereby giving a justification for the proposed estimator (4.3).

It may also be of interest to estimate the integrated hazard for an individual with a specified covariate $\mathbf{Z}_0$ fixed over time. This is $A(t;\mathbf{Z}_0) = \exp(\beta_0^{\mathsf{T}}\mathbf{Z}_0)A_0(t)$ and it is estimated by

$$\widehat{A}(t;\mathbf{Z}_0) = \exp(\widehat{\beta}^{\mathsf{T}}\mathbf{Z}_0)\widehat{A}_0(t;\widehat{\beta}). \tag{4.7}$$

In Theorem 5 in Section 6 we will show that, as $n \to \infty$,

$$\sqrt{n}(\widehat{A}(\cdot\,;\mathbf{Z}_0) - A(\cdot\,;\mathbf{Z}_0)) \tag{4.8}$$

converges weakly to a mean zero Gaussian process with a covariance function which may be estimated uniformly consistently by $n\hat{\sigma}^2(s,t)$ where

$$\hat{\sigma}^2(s,t) = \left\{\exp(\hat{\beta}^{\mathsf{T}}\mathbf{Z}_0)\right\}^2 \tag{4.9}$$
$$\times \left\{\hat{\omega}^2(s\wedge t;\hat{\beta})) + \left(\hat{\mathbf{B}}(s;\hat{\beta}) - \hat{A}(s;\hat{\beta})\mathbf{Z}_0\right)^{\mathsf{T}} \mathcal{I}(\hat{\beta})^{-1} \left(\hat{\mathbf{B}}(t;\hat{\beta}) - \hat{A}(t;\hat{\beta})\mathbf{Z}_0\right)\right\},$$

with

$$\hat{\omega}^2(t;\beta) = \sum_{t_j\le t} \frac{1}{\left\{\sum_{l\in\tilde{\mathcal{R}}(t_j)} Y_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(t_j))/p_{t_j}(l)\right\}^2}, \tag{4.10}$$

and

$$\hat{\mathbf{B}}(t;\beta) = \sum_{t_j\le t} \frac{\sum_{l\in\tilde{\mathcal{R}}(t_j)} Y_l(t_j)\mathbf{Z}_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(t_j))/p_{t_j}(l)}{\left\{\sum_{l\in\tilde{\mathcal{R}}(t_j)} Y_l(t_j)\exp(\beta^{\mathsf{T}}\mathbf{Z}_l(t_j))/p_{t_j}(l)\right\}^2}. \tag{4.11}$$

The results for the cumulative baseline relative hazard estimator (4.3) are obtained by inserting $\mathbf{Z}_0 = 0$ above.

Using product-integral notation (Andersen *et al.*, 1992; Section II.6), the survival probability

$$S(t;\mathbf{Z}_0) = \prod_{u\le t}(1 - \mathrm{d}A(u;\mathbf{Z}_0)) = \exp\left(-A(t;\mathbf{Z}_0)\right)$$

for an individual with a fixed covariate value $\mathbf{Z}_0$ can be estimated by

$$\begin{aligned}\hat{S}(t;\mathbf{Z}_0) &= \prod_{u\le t}\left(1 - \mathrm{d}\hat{A}(u;\mathbf{Z}_0)\right)\\ &= \prod_{t_j\le t}\left(1 - \frac{\exp(\hat{\beta}^{\mathsf{T}}\mathbf{Z}_0)}{\sum_{l\in\tilde{R}_j} Y_l(t_j)\exp(\hat{\beta}^{\mathsf{T}}\mathbf{Z}_l(t_j))/p_{t_j}(l)}\right).\end{aligned} \tag{4.12}$$

By the argument in Section VII.2.3 in Andersen *et al.* (1992)

$$\sqrt{n}(\hat{S}(\cdot\,;\mathbf{Z}_0) - S(\cdot\,;\mathbf{Z}_0)) \tag{4.13}$$

is asymptotically equivalent to $-S(\cdot\,;\mathbf{Z}_0)$ times (4.8), and it follows that, as $n\to\infty$, (4.13) converges weakly to a mean zero Gaussian process with a covariance function which may be estimated uniformly consistently by $n\hat{S}(s;\mathbf{Z}_0)\hat{S}(t;\mathbf{Z}_0)\hat{\sigma}^2(s,t)$.

5 Examples of specific sampling schemes

The methodology we have presented in the preceding sections provides analytic tools for a very large class of sampling schemes. We illustrate its flexibility by deriving the partial likelihood for a few diverse designs and also derive the estimator (4.3) for the integrated baseline hazard for a few of them. In each situation considered, the sometimes complex $\pi_t(\mathbf{r}|i)$ in the partial likelihood (3.3) are not needed for the actual analysis of the data. Cancellation of common terms or multiplication by quantities which do not depend on β leads to considerable reduction yielding a partial likelihood of the form of (1.4).

As we will discuss in more detail in Section 10, that while the partial likelihood provides a method of valid estimation of the regression parameters, there is no guarantee that the estimator $\widehat{\beta}$ will be efficient relative to "the best" method of analysis. This is illustrated in Examples 5.5 and 5.6.

Remember in what follows that the sampling distributions $\pi_t(\cdot \mid i)$ are defined over sets $\mathbf{r}$ in $\mathcal{P}_i$ so that that $\pi(\mathbf{r}|i) = 0$ of $i \notin \mathbf{r}$. We let $|\mathbf{r}|$ denote the number of elements in the set $\mathbf{r}$.

Example 5.1 Full cohort. The full cohort partial likelihood is a special case in which the entire risk set $\mathcal{R}(t) = \{i : Y_i(t) = 1\}$ is sampled with probability one. In our notation then, $\pi_t(\mathbf{r}|i) = I(\mathbf{r} = \mathcal{R}(t))$ for all $i \in \mathcal{R}(t)$, and the usual Cox partial likelihood for the full data set is recovered. Noting that (4.1) and (4.2) are fulfilled with $\pi_t(\mathcal{R}(t)) = p_t(i) \equiv 1$, $A_0(t;\widehat{\beta})$ is as in (1.3), the usual Breslow estimator of the integrated baseline hazard function.

Example 5.2 Nested case-control sampling. The most common type of cohort sampling technique is nested case-control sampling, in which $m - 1$ controls are randomly sampled, without replacement from those at risk at the failure's failure time. Here we assume that $m > 1$ is fixed. Letting $n(t) = \sum_{i=1}^{n} Y_i(t) = |\mathcal{R}(t)|$ denote the number at risk at time t, this sampling scheme is specified by

$$\pi_t(\mathbf{r}|i) = \binom{n(t)-1}{m-1}^{-1} I(\mathbf{r} \in \mathcal{P}_i, \mathbf{r} \subset \mathcal{R}(t), |\mathbf{r}| = m),$$

which is the same for each $i \in \mathbf{r}$ and thus drops out of (3.3) leaving the usual partial likelihood (Oakes, 1981) for nested case-control sampling. Further, (4.1) and (4.2) are satisfied with

$$\pi_t(\mathbf{r}) = \binom{n(t)}{m}^{-1} I(\mathbf{r} \subset \mathcal{R}(t), |\mathbf{r}| = m) \quad \text{and} \quad p_t(i) = \frac{m}{n(t)}$$

and, from (4.3), the estimator of the cumulative hazard function is

$$\widehat{A}_0(t;\widehat{\beta}) = \sum_{t_j \le t} \frac{1}{\sum_{l \in \widetilde{\mathcal{R}}_j} \exp(\widehat{\beta}^{\mathsf{T}} \mathbf{Z}_l(t_j)) n(t_j)/m}$$

with variance estimator from (4.9).

Example 5.3 Stratified nested case-control sampling. In this extension of nested case-control sampling (Langholz and Borgan, 1992), control sampling is performed within sampling strata. In general, let $C_i(t)$ be $(\mathcal{F}_t)$-predictable sampling strata indicators with

$C_i(t) \in \mathcal{C}$, a (small) finite set of indices. Define $\mathcal{R}_l(t) = \{i : Y_i(t) = 1, C_i(t) = l\}$ to be sampling stratum l with $n_l(t) = |\mathcal{R}_l(t)|$. If a subject, say i, fails at time t, then m_l controls are randomly sampled without replacement from $\mathcal{R}_l(t)$ except for the failure's stratum $\mathcal{R}_{C_i(t)}(t)$ from which $m_{C_i(t)} - 1$ are sampled from the $n_{C_i(t)}(t) - 1$ non-failures. As a technicality, which we do not consider further and which causes no difficulties, the number of controls could also depend on time. Specifically, if $n_l(t) < m_l$ all subjects from $\mathcal{R}_l(t)$ would be included in the sample. The probability structure for this sampling scheme is given by

$$\pi_t(\mathbf{r}|i) = \left[\prod_{l \in \mathcal{C}} \binom{n_l(t)}{m_l}\right]^{-1} \frac{n_{C_i(t)}(t)}{m_{C_i(t)}} I(\mathbf{r} \in \mathcal{P}_i, \mathbf{r} \subset \mathcal{R}(t), |\mathbf{r} \cap \mathcal{R}_l(t)| = m_l;\ l \in \mathcal{C}),$$

which yields weights

$$w_i(t) = \frac{n_{C_i(t)}(t)}{m_{C_i(t)}} \tag{5.1}$$

so that in the partial likelihood (1.4), the relative risk for a subject from a given stratum is weighted by the inverse of the proportion of the stratum sampled. Note that the case contribution is weighted no differently than the controls.

The sampling probabilities simplify into the form of (4.1) and (4.2) with

$$\pi_t(\mathbf{r}) = \left[\prod_{l \in \mathcal{C}} \binom{n_l(t)}{m_l}\right]^{-1} I(\mathbf{r} \subset \mathcal{R}(t), |\mathbf{r} \cap \mathcal{R}_l(t)| = m_l;\ l \in \mathcal{C})$$

and $p_t(i) = w_i(t)^{-1}$. Thus, the baseline hazard estimator becomes

$$\widehat{A}_0(t; \widehat{\beta}) = \sum_{t_j \le t} \frac{1}{\sum_{l \in \widetilde{\mathcal{R}}_j} \exp(\widehat{\beta}^\top \mathbf{Z}_l(t_j)) w_l(t_j)}$$

with variance estimator given by (4.9).

Because $\mathcal{F}_t$ contains failure, censoring, covariate, and sampling histories up to time t, the sampling strata may be defined in some quite diverse ways. We describe some interesting designs from this class.

Stratification based on absolute exposure status. This design is discussed in Langholz and Borgan (1992). For simplicity consider a dichotomous variable $Z_1(t) = 0$ or 1 indicating unexposed or exposed. Assume that $Z_1(t)$ is known for all subjects at risk at t (cf. the comment at the end of Section 2), and define the sampling strata as $\mathcal{R}_l(t) = \{i : i \in \mathcal{R}(t), Z_{i1}(t) = l\}$ for $l = 0, 1$. The $C_i(t)$ are simply defined to be $Z_{i1}(t)$. When a failure occurs, m_0 and m_1 subjects are "sampled" (with the understanding that the failure is included in the sample) from the unexposed, $\mathcal{R}_0(t)$, and exposed, $\mathcal{R}_1(t)$, at risk subjects, respectively. Additional covariate information would then be collected on this stratified sample.

Stratification based on relative exposure status. Now consider a multivalued or continuous $Z_1(t)$ and assume that this is known for all individuals at risk at time t. Sampling

strata may then be defined in terms of the distribution of the $Z_{i1}(t)$ for the at risk subjects. For example, define the sampling strata at t as some empirical quantile intervals of the $\{Z_{i1}(t) : i \in \mathcal{R}(t)\}$ and let $C_i(t)$ be the particular interval that $Z_{i1}(t)$ falls into.

Stratification based on sampling history. As a specific example of this situation, we give a technique for ensuring that each sampled risk set adds m_1 new subjects to the sample. Consider sampling m_0 from the set of those not sampled in any previous risk set (stratum 0) and m_1 from those who have been sampled (stratum 1). Let $\mathcal{S}(t) = \mathcal{R}(t) \setminus \left(\cup_{t_j<t}\tilde{\mathcal{R}}_j\right)$, the set of those not yet picked in any sampled risk set. Then (5.1) are the appropriate weights with $n_0(t) = |\mathcal{S}(t)|$ and $n_1(t) = n(t) - n_0 > (t)$.

Example 5.4 Quota sampling: negative hypergeometric sampling. Consider processes $C_i(t) \in \{0,1\}$ with a role similar to that in stratified sampling discussed in the previous example and let m_l, $n_l(t)$, and $\mathcal{R}_l(t)$, $l = 0,1$ be as defined there. In this sampling method, if a subject i fails at time t, controls are sampled sequentially until m_1 subjects are selected from $\mathcal{R}_1(t)$. As before, if $i \in \mathcal{R}_1(t)$, the failure i is counted as one of the m_1. Such a sampling scheme might be considered when a (dichotomous) exposure of interest is somewhat rare and is expensive to determine but obtaining the additional covariate information needed to perform a complete analysis is inexpensive. Thus, once exposure status for a subject is determined there is no reason not to include the subject as a control in the sampled risk set. Further, the intuitive idea that the amount of "exposure discordance" within the sampled risk sets determines the efficiency of the sample suggests that it may be possible to ensure that a given level of (relative) efficiency for estimating the effect of exposure is achieved by specifying that a fixed number of exposed subjects be in the sampled risk set.

The probability of sampling a particular set depends on the size of the set and whether $C_i(t)$ for the failure is 1 or 2. Specifically, the size of the sampled risk set has a negative hypergeometric distribution (Schuster and Sype, 1987) with

$$\pi_t(\mathbf{r}|i) = \begin{cases} \frac{\binom{|\mathbf{r}|-2}{m_1-2}\binom{n(t)-|\mathbf{r}|}{n_1(t)-m_1}}{\binom{n(t)-1}{n_1(t)-1}} I(|\mathbf{r} \cap \mathcal{R}_1(t)| = m_1) & \text{if } C_i(t) = 1 \\ \frac{\binom{|\mathbf{r}|-2}{m_1-1}\binom{n(t)-|\mathbf{r}|}{n_1(t)-m_1}}{\binom{n(t)-1}{n_1(t)}} I(|\mathbf{r} \cap \mathcal{R}_1(t)| = m_1) & \text{if } C_i(t) = 0 \end{cases}$$

Cancellation of common factors yields considerable reduction leaving weights $w_i(t) = (m_1 - 1)/n_1(t)$ or $m_0/n_0(t)$ if $C_i(t) = 1$ or 0, respectively, to be included in the partial likelihood (1.4).

Note that if $m_1 = 1$, it is not possible to estimate the regression parameters since all exposed subjects, i.e., those with $C_i(t) = 1$, are weighted by zero. This is because if the failure is exposed, the sampled risk set consists only of that failure making estimation impossible. One possible solution is to (simple) randomly sample one control, without regard to the exposure status of the failure, before starting the quota sampling. This would assure that exposed failures are almost always matched to an unexposed control (since exposure is assumed rare) and that there would be one (or rarely two) exposed controls if the failure is unexposed.

In the situations we are suggesting that this design might be useful, the $n_l(t)$ will not be known. One possible strategy is to replace the $n_l(t)$ by method of moments estimators. Then $\hat{n}_1(t) = n(t)(m_1 - 1)/(|\mathbf{r}| - 1)$ or $n(t)\ m_1/(|\mathbf{r}| - 1) - 1$ if $C_i(t) = 1$ or 0, respectively,

and $\hat{n}_0(t) = n(t) - \hat{n}_1(t)$. The validity of such an approach and the proper adjustments to the variance when the weights are estimated require further study.

Not surprisingly, the negative binomial distribution with probability parameter $n_1(t)/n(t)$ yields identical weights to those given in the preceding paragraph. Using the negative multinomial distribution, this leads us to conjecture that the extension to $C_i(t) \in \{0, \ldots, L\}$ and m_L fixed will yield weights $w_L(t) = (m_L - 1)/n_L$ and $w_i(t) = m_l/n_l(t)$ if $C_i(t) = l \neq L$.

Example 5.5 Nested case-control sampling with variable matching ratio. Instead of fixed m, consider using a variable matching ratio, possibly depending upon characteristics of the case. Specifically, in this class of sampling designs first the size of the sampled risk set is (randomly) determined and then a simple nested case-control sample of this size is selected. For instance, suppose that exposure status ($Z_1(t) = 0$ or 1 indicating unexposed or exposed) is to be gathered on a sample of the cohort. One might first ascertain the case's exposure status, then, based on that status, decide the number of controls to be sampled. We conjecture that when there are many cases and exposure is rare, matching more controls to exposed cases is more efficient than fixing the matching ratio for all sampled risk sets.

Let $\tilde{n}(t)$ be the size of the sampled risk set if there is a failure at time t. We assume that $\tilde{n}(t)$ is random with a (predictable) probability distribution on $\{1, \ldots, n(t)\}$ which may depend on who failed at that time. We may then specify the sampling probabilities as functions of the size of $\mathbf{r}$ with

$$\pi_t(\mathbf{r}|i) = \binom{n(t)-1}{|\mathbf{r}|-1}^{-1} I(\mathbf{r} \in \mathcal{P}_i, \mathbf{r} \subset \mathcal{R}(t)) \mathrm{P}(\tilde{n}(t) = |\mathbf{r}| \mid i, \mathcal{F}_{t-}).$$

Because the binomial coefficient is common to all $i \in \mathbf{r}$ cancellation yields weights $w_i(t) = \mathrm{P}(\tilde{n}(t) = |\mathbf{r}| \mid i, \mathcal{F}_{t-})$ in (1.4).

Now if we adopt the variable matching scheme as implied above, in which with probability 1, exactly m_1 or m_2 controls would be picked if the case is unexposed or exposed, respectively, for $i \in \mathbf{r}$, $P(\tilde{n}(t) = |\mathbf{r}| \mid i, \mathcal{F}_{t-})$ is one if subject i has the same exposure status as the case and zero if they differ. This results in sampled risk sets matched for exposure status making it impossible to estimate the effect of exposure. Whether this is a property inherent in the design, i.e., there is no method of analysis which yields a consistent estimator, or the partial likelihood has zero efficiency, is not known.

An approach which leads to a more reasonable analysis is to consider two different sample sizes, say m_1 and m_2, with $P(\tilde{n}(t) = m_1 | i, \mathcal{F}_{t-}) = \pi_{Z_{i1}(t)}$ and $P(\tilde{n}(t) = m_2 | i, \mathcal{F}_{t-}) = 1 - \pi_{Z_{i1}(t)}$ for some fixed probabilities π_0 and π_1. One might also consider $\tilde{n}(t)$ as binomial with probabilities π_0 and π_1 for i unexposed and exposed, respectively.

Example 5.6 Case-cohort sampling. Prentice (1986*a*) presents case-cohort sampling in which a subcohort $\tilde{C}$ is randomly sampled from the full cohort at $t = 0$. He shows heuristically that the partial likelihood for this design does not make use of the non-subcohort failures in the estimation of β and proposes a "pseudo-likelihood" approach. In our formulation, since $\tilde{C} \in \mathcal{F}_0$,

$$\pi_t(\mathbf{r}|i) = I\left(\mathbf{r} = (\tilde{C} \cap \mathcal{R}(t)) \cup \{i\}\right).$$

Thus, if $i \in \tilde{C}$ fails at t_j then $\tilde{\mathcal{R}}_j = \tilde{C} \cap \mathcal{R}(t_j)$ and $\pi_t(\tilde{\mathcal{R}}_j | l) = 1$ for all $l \in \tilde{\mathcal{R}}_j$, but if $i \notin \tilde{C}$, $\tilde{\mathcal{R}}_j = \left(\tilde{C} \cap \mathcal{R}(t_j)\right) \cup \{i\} \neq \tilde{C} \cap \mathcal{R}(t_j)$ and $\pi_t(\tilde{\mathcal{R}}_j | l) = I(l = i)$ since this sampled

risk set would occur with probability zero if a subcohort member failed. Thus, if i is a non-subcohort failure, the partial likelihood weights subcohort members by zero, leaving a contribution of one for that sampled risk set, confirming Prentice's conclusion. This is an example where the partial likelihood (3.3) is clearly inefficient for the design.

6 Asymptotic properties of the estimators

In this section we study the large sample properties of the estimator $\widehat{\beta}$ for the regression parameters and the estimator $\widehat{A}_0(t,\widehat{\beta})$ for the integrated baseline hazard. We therefore consider a sequence of models indexed by n of the form defined in Section 2 with processes $N_{(i,\mathbf{r})}^{(n)}$, $Y_i^{(n)}$, $\mathbf{Z}_i^{(n)}$, $\pi_{\cdot}^{(n)}(\mathbf{r} \mid i)$, etc. depending on n. For ease of notation we will, however, drop the superscript (n) from the notation, but the reader should keep in mind that these quantities depend on n whereas the true parameter values β_0 and α_0 are the same for all n.

We remind the readers of the definitions (3.4) - (3.6) and write $C_t(\beta)$, $\mathbf{U}_t(\beta)$ and $\mathcal{I}_t(\beta)$ for (3.7) - (3.9), respectively, when the integral is taken over $[0,t]$ instead of $[0,\tau]$. Moreover we denote the jth component of the vector $\mathbf{U}_t(\beta)$ by $U_t^j(\beta)$ and the (j,k)th element of the matrix $\mathcal{I}_t(\beta)$ by $\mathcal{I}_t^{jk}(\beta)$.

For the proofs we will need the following conditions, where the norm of a vector $\mathbf{a} = (a_i)$ or a matrix $\mathbf{A} = \{a_{ij}\}$ is $\|\mathbf{a}\| = \sup_i |a_i|$ and $\|\mathbf{A}\| = \sup_{i,j} |a_{ij}|$, respectively. Conditions 1 and 2 are assumptions; both are assumed to hold in what follows. In Section 7 we show how the remaining conditions are satisfied for some specific model assumptions and sampling schemes.

Condition 1 $A_0(\tau) < \infty$.

Condition 2 *The covariate processes $\mathbf{Z}_i(t), i = 1,2,\ldots,n$ are left continuous and $(\mathcal{F}_t)$-adapted.*

Condition 3 *For $(\rho,\gamma) = (0,2)$ and $(\rho,\gamma) = (2,0)$ define*

$$Q^{(\rho,\gamma)}(\beta_0,t) = \frac{1}{n}\sum_{\mathbf{r}\in\mathcal{P}} \mathbf{E}_{\mathbf{r}}(\beta_0,t)^{\otimes\rho} S_{\mathbf{r}}^{(\gamma)}(\beta_0,t), \tag{6.1}$$

and assume that there exists functions $q^{(\rho,\gamma)}$ such that for all $t \in [0,\tau]$,

$$Q^{(\rho,\gamma)}(\beta_0,t) \xrightarrow{\mathrm{P}} q^{(\rho,\gamma)}(\beta_0,t) \quad \text{as } n \to \infty. \tag{6.2}$$

Condition 4 *The $p \times p$-matrix matrix $\mathbf{\Sigma} = \{\sigma_{jk}\}$ given by*

$$\mathbf{\Sigma} = \int_0^\tau [q^{(0,2)}(\beta_0,t) - q^{(2,0)}(\beta_0,t)]\alpha_0(t)\mathrm{d}t$$

is positive definite.

Condition 5 *For any n and each $\mathbf{r} \in \mathcal{P}$ there exists a locally bounded predictable process $X_{\mathbf{r},n}$ not depending on β such that for all $t \in [0,\tau]$*

$$\|\mathbf{Z}_i(t)\| \leq X_{\mathbf{r},n}(t) \quad \textit{for all } i \in \mathbf{r}. \tag{6.3}$$

Moreover there exists a $b_0 > 3\|\beta_0\|$ such that, with

$$D_n(t) = \frac{1}{n}\sum_{\mathbf{r}\in\mathcal{P}} \exp(b_0 X_{\mathbf{r},n}(t)) \sum_{i\in\mathbf{r}} \pi_t(\mathbf{r} \mid i),$$

there exists $D(t)$ such that

$$D_n(t) \xrightarrow{\mathrm{P}} D(t) \quad \textit{and} \quad \int_0^\tau D_n(t)\alpha_0(t)\mathrm{d}t \xrightarrow{\mathrm{P}} \int_0^\tau D(t)\alpha_0(t)\mathrm{d}t < \infty, \tag{6.4}$$

as $n \to \infty$.

Before we derive the asymptotic properties of the maximum partial likelihood estimator $\widehat{\beta}$, we state some useful consequences of these conditions.

Let $\mathcal{B}_0$ be an open neighborhood of β_0 with $\sup\{\|\beta\| : \beta \in \mathcal{B}_0\} \leq b_0/3$. Then by (3.4) and (6.3) we have for $\beta \in \mathcal{B}_0$

$$\begin{aligned} & \exp(-(b_0/3)X_{\mathbf{r},n}(t)) \sum_{i\in\mathbf{r}} \pi_t(\mathbf{r} \mid i) \\ \leq\ & S_{\mathbf{r}}^{(0)}(\beta,t) \leq \\ & \exp((b_0/3)X_{\mathbf{r},n}(t)) \sum_{i\in\mathbf{r}} \pi_t(\mathbf{r} \mid i). \end{aligned} \tag{6.5}$$

Moreover, for $\gamma = 1, 2$,

$$\|S_{\mathbf{r}}^{(\gamma)}(\beta,t)\| \leq X_{\mathbf{r},n}(t)^\gamma S_{\mathbf{r}}^{(0)}(\beta,t), \tag{6.6}$$

and by (3.5), (3.6) and (6.1)

$$\|\mathbf{E}_{\mathbf{r}}(\beta,t)\| \leq X_{\mathbf{r},n}(t), \tag{6.7}$$

$$\|\mathbf{V}_{\mathbf{r}}(\beta,t)\| \leq X_{\mathbf{r},n}(t)^2 \tag{6.8}$$

and

$$\|Q^{(\rho,\gamma)}(\beta_0,t)\| \leq \frac{1}{n}\sum_{\mathbf{r}\in\mathcal{P}} X_{\mathbf{r},n}(t)^{\rho+\gamma} S_{\mathbf{r}}^{(0)}(\beta_0,t). \tag{6.9}$$

To show the convergence in probability for various integrals that arise below, we use a version of a dominated convergence of Hjort and Pollard (Hjort and Pollard, 1993):

Proposition 1 *Suppose $A_0(\tau) < \infty$ and let $0 \leq X_n(s) \leq D_n(s)$ be left-continuous random processes on the interval $[0,\tau]$. Suppose $D_n(s) \xrightarrow{\mathrm{P}} D(s)$ and $X_n(s) \xrightarrow{\mathrm{P}} X(s)$ for almost all s and that $\int_0^\tau D_n(s)\alpha_0(s)\mathrm{d}s \xrightarrow{\mathrm{P}} \int_0^\tau D(s)\alpha_0(s)\mathrm{d}s < \infty$. Then $\int_0^t X_n(s)\alpha_0(s)\mathrm{d}s \xrightarrow{\mathrm{P}} \int_0^t X(s)\alpha_0(s)\mathrm{d}s$ for all $t \in [0,\tau]$.*

By (6.5), the right hand side of (6.9) is bounded by a constant times D_n, and it follows by the above dominated convergence theorem and Condition 5 that, for all $t \in [0, \tau]$,

$$\int_0^t Q^{(\rho,\tau)}(\beta_0, u)\alpha_0(u)\mathrm{d}u \xrightarrow{\mathrm{P}} \int_0^t q^{(\rho,\tau)}(\beta_0, u)\alpha_0(u)\mathrm{d}u \tag{6.10}$$

as $n \to \infty$. In particular, by (3.6), (3.11) and Condition 4,

$$n^{-1}\langle \mathbf{U}_\cdot(\beta_0)\rangle(\tau) = \frac{1}{n}\int_0^\tau \sum_{r\in\mathcal{P}} \mathbf{V}_r(\beta_0, u)S_r^{(0)}(\beta_0, u)\alpha_0(u)\mathrm{d}u \xrightarrow{\mathrm{P}} \mathbf{\Sigma}. \tag{6.11}$$

We now prove that the estimator $\widehat{\beta}$ for the regression parameters is consistent.

Theorem 1 *Assume Conditions 1-5. Then the estimator $\widehat{\beta}$ maximizing* (3.3) *is consistent for β_0.*

Proof: By a Taylor series expansion we have for any $\beta = (\beta_1, \ldots, \beta_p)^\mathsf{T} \in \mathcal{B}_0$

$$U_\tau^j(\beta) = U_\tau^j(\beta_0) - \sum_{k=1}^p (\beta_k - \beta_{k0})\mathcal{I}_\tau^{jk}(\beta_0) + \frac{1}{2}\sum_{k=1}^p\sum_{l=1}^p (\beta_k - \beta_{k0})(\beta_l - \beta_{l0})R_\tau^{jkl}(\beta^*),$$

where β^* is on the line segment joining β and β_0, and

$$R_\tau^{jkl}(\beta) = \frac{\partial^3 C_\tau(\beta)}{\partial\beta_j\partial\beta_k\partial\beta_l}. \tag{6.12}$$

We need to show that for all j, k, l

$$n^{-1}U_\tau^j(\beta_0) \xrightarrow{\mathrm{P}} 0, \tag{6.13}$$

and

$$n^{-1}\mathcal{I}_\tau^{jk}(\beta_0) \xrightarrow{\mathrm{P}} \sigma_{jk}, \tag{6.14}$$

as $n \to \infty$, and that there exists a finite constant K such that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|n^{-1}R_\tau^{jkl}(\beta)\right| \le K \text{ for all } j, k, l, \text{ and all } \beta \in \mathcal{B}_0\right) = 1. \tag{6.15}$$

For from (6.13) – (6.15) it follows e.g. by the argument in Billingsley (1961, pp. 12-13) that with probability tending to one there exists a consistent solution to the score equations $\mathbf{U}_\tau(\beta) = \mathbf{0}$. But from this the theorem follows since $C_\tau(\beta)$ is concave, and hence the score equations have at most one solution.

To prove (6.13) we use Lenglart's inequality (e.g. Andersen *et al.*, 1992, Section II.5.2) to get for all $\delta, \eta > 0$

$$\mathrm{P}\left(\sup_{t\in[0,\tau]}\left|n^{-1}U_t^j(\beta_0)\right| \ge \eta\right) \le \frac{\delta}{\eta^2} + \mathrm{P}\left(n^{-2}\langle U^j(\beta_0)\rangle(\tau) \ge \delta\right).$$

But by (6.11) $n^{-1}\langle U^j(\beta_0)\rangle(\tau) \xrightarrow{\mathrm{P}} \sigma_{jj} < \infty$, and (6.13) follows.

To prove (6.14), we write $V_r^{jk}(\beta, t)$ for the (j, k)th element of (3.6). Then by (2.13), (2.15), (3.4) and Lenglart's inequality we have for all $\delta, \eta > 0$

$$\mathrm{P}\left(\sup_{t\in[0,\tau]}\left|n^{-1}\int_0^t\sum_{r\in\mathcal{P}}V_r^{jk}(\beta_0,u)\mathrm{d}M_r(u)\right|\geq\eta\right) \tag{6.16}$$
$$\leq \frac{\delta}{\eta^2}+\mathrm{P}\left(n^{-2}\int_0^\tau\sum_{r\in\mathcal{P}}\left(V_r^{jk}(\beta_0,u)\right)^2 S_r^{(0)}(\beta_0,u)\alpha_0(u)\mathrm{d}u\geq\delta\right).$$

But by (6.8)

$$n^{-1}\int_0^\tau\sum_{r\in\mathcal{P}}\left(V_r^{jk}(\beta_0,u)\right)^2 S_r^{(0)}(\beta_0,u)\alpha_0(u)\mathrm{d}u$$
$$\leq n^{-1}\int_0^\tau\sum_{r\in\mathcal{P}}X_{r,n}(u)^4 S_r^{(0)}(\beta_0,u)\alpha_0(u)\mathrm{d}u,$$

which by (6.5) is bounded by (6.4) times a constant. Thus, by Condition 5 and (6.16), $1/n$ times the second term on the right hand side of (3.12) converges to zero in probability, and (6.14) follows by (3.12) and (6.11).

Finally, to prove (6.15), we first note that (3.7) and (6.12) give, for any n and all j,k,l and $\beta\in\mathcal{B}_0$,

$$\left|n^{-1}R_\tau^{jkl}(\beta)\right|\leq\frac{6}{n}\int_0^\tau X_{r,n}(u)^3\mathrm{d}N_r(u).$$

By another use of Lenglart's inequality and (2.13), (2.14) and (3.4) we have for any $C,K>0$

$$\mathrm{P}\left(\frac{6}{n}\int_0^\tau\sum_{r\in\mathcal{P}}X_r(u)^3\mathrm{d}N_r(u)\geq K\right) \tag{6.17}$$
$$\leq \frac{C}{K}+\mathrm{P}\left(\frac{6}{n}\int_0^\tau\sum_{r\in\mathcal{P}}X_r(u)^3 S_r^{(0)}(\beta_0,u)\alpha_0(u)\mathrm{d}u\geq C\right).$$

By (6.5) and Condition 5 it is seen that the second term on the right hand side tends to zero as $n\to\infty$ if C is chosen large enough. Thus the right hand side can be made arbitrarily small for n and K large enough, and (6.15) is proved. □.

Before we study the asymptotic distribution of $\widehat{\beta}$, we give a consistent estimator of Σ, the inverse of the asymptotic covariance matrix.

Theorem 2 *Assume Conditions 1-5. Then for any $\beta^*\xrightarrow{\mathrm{P}}\beta_0$ we have*

$$\frac{1}{n}\mathcal{I}_\tau(\beta^*)\xrightarrow{\mathrm{P}}\Sigma,$$

as $n\to\infty$, where Σ is defined in Condition 4.

Proof: By a Taylor series expansion we have when $\beta^* \in \mathcal{B}_0$

$$n^{-1}\mathcal{I}_\tau^{jk}(\beta^*) = n^{-1}\mathcal{I}_\tau^{jk}(\beta_0) - n^{-1}\sum_{l=1}^{p}(\beta_l^* - \beta_{l0})R_\tau^{jkl}(\check{\beta}),$$

where $R_\tau^{jkl}(\beta)$ is defined by (6.12) and $\check{\beta}$ is on the line segment joining β^* and β_0. By (6.14), the first term converges in probability to σ_{jk} as $n \to \infty$. Moreover, by (6.15), the second term is bounded in probability by $pK\|\beta^* - \beta_0\|$ for some finite constant K not depending on β^*, and the theorem is proved. □

We now demonstrate the asymptotic normality of the maximum partial likelihood estimator.

Theorem 3 *Assume Conditions 1-5 and let $\widehat{\beta}$ be the estimator maximizing* (3.3). *Then*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^{-1}),$$

as $n \to \infty$, where Σ is defined in Condition 4.

Proof: Taylor expanding $U_\tau^j(\widehat{\beta})$ around β_0 gives when $\widehat{\beta} \in \mathcal{B}_0$

$$0 = n^{-1/2}U_\tau^j(\widehat{\beta}) = n^{-1/2}U_\tau^j(\beta_0) - \sum_{k=1}^{p}\sqrt{n}(\widehat{\beta}_k - \beta_{k0})\, n^{-1}\mathcal{I}_\tau^{jk}(\beta^*), \tag{6.18}$$

where β^* is on the line segment between $\widehat{\beta}$ and β_0. Hence, using Theorems 1 and 2, it sufficient to prove that

$$n^{-1/2}\mathbf{U}_\tau(\beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma) \tag{6.19}$$

as $n \to \infty$.

To this end we apply Rebolledo's martingale central limit theorem as presented e.g. in Section II.5.1 in Andersen *et al.* (1992). By (6.11) the (vector valued) local square integrable martingale $n^{-1/2}\mathbf{U}_\cdot(\beta_0)$ has a predictable variation process which evaluated at τ converges in probability to Σ. Thus condition (2.5.1) of Theorem II.5.1 in Andersen *et al.* (1992) is fulfilled for $t = \tau$. To prove the Lindeberg condition (specified by (2.5.3) and (2.5.8) in Andersen *et al.*, 1992) we introduce $E_{\mathbf{r}}^j(\beta, t)$ for the j th component of (3.5). Then by a Chebychev type inequality we have for all j any $\epsilon > 0$

$$\begin{aligned}
&\frac{1}{n}\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}}\left\{Z_{ij}(u) - E_{\mathbf{r}}^j(\beta_0, u)\right\}^2 I\left\{n^{-1/2}\left|Z_{ij}(u) - E_{\mathbf{r}}^j(\beta_0, u)\right| > \epsilon\right\}\lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \\
&\quad\le \frac{1}{\epsilon n^{3/2}}\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}}\left|Z_{ij}(u) - E_{\mathbf{r}}^j(\beta_0, u)\right|^3\lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \\
&\quad\le \frac{8}{\epsilon n^{3/2}}\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} X_{\mathbf{r},n}(u)^3 S_{\mathbf{r}}^{(0)}(\beta_0, u)\alpha_0(u)\mathrm{d}u,
\end{aligned}$$

where the last inequality follows by (2.13), (3.4), (6.3) and (6.7). But by (6.5) and Condition 5 the right hand side tends to zero in probability as $n \to \infty$, and the Lindeberg

condition is proved. □

We then turn to a study of the large sample properties of the estimator $\widehat{A}_0(t;\widehat{\beta})$ for the baseline hazard. We then assume that the sampling distributions fulfill (4.1) and (4.2) and impose the following extra conditions.

Condition 6 *For $J(t)$ given by (4.4) we have*

$$\inf_{t\in[0,\tau]} J(t) \xrightarrow{\mathrm{P}} 1.$$

Condition 7 *Let $X_{\mathbf{r},n}$ and b_0 be given by Condition 5. Then for $\gamma = 0,1,2$*

$$n^\gamma \int_0^\tau J(u) \sum_{\mathbf{r}\in\mathcal{P}} \pi_u(\mathbf{r})^{\gamma+1} \exp\left(b_0 X_{\mathbf{r},n}(u)\right) \left\{\sum_{i\in\mathbf{r}} \pi_u(\mathbf{r}\,|\,i)\right\}^{-\gamma} \alpha_0(u)\mathrm{d}u \tag{6.20}$$

all converge in probability to finite quantities as $n \to \infty$.

Condition 8 *There exist functions e and ϕ such that for all $t \in [0,\tau]$*

$$J(t) \sum_{\mathbf{r}\in\mathcal{P}} \pi_t(\mathbf{r}) \mathbf{E}_{\mathbf{r}}(\beta_0, t) \xrightarrow{\mathrm{P}} e(\beta_0, t), \tag{6.21}$$

and

$$nJ(t) \sum_{\mathbf{r}\in\mathcal{P}} \pi_t(\mathbf{r})^2 \left\{S_{\mathbf{r}}^{(0)}(\beta_0, t)\right\}^{-1} \xrightarrow{\mathrm{P}} \phi(\beta_0, t), \tag{6.22}$$

as $n \to \infty$.

We first prove the following lemma.

Lemma 1 *Let $\widehat{\mathbf{B}}(t;\beta)$ be given by (4.11), and assume that Conditions 1-8 hold. Then for any $\beta^* \xrightarrow{\mathrm{P}} \beta_0$ we have*

$$\sup_{t\in[0,\tau]} \left\| \widehat{\mathbf{B}}(t;\beta^*) - \mathbf{B}(t;\beta_0) \right\| \xrightarrow{\mathrm{P}} 0,$$

as $n \to \infty$, with

$$\mathbf{B}(t;\beta_0) = \int_0^t e(\beta_0, u)\alpha_0(u)\mathrm{d}u, \tag{6.23}$$

and $e(\beta_0, u)$ defined in (6.21).

Proof: First note that by (3.4), (3.5) and (4.1) we may write

$$\widehat{\mathbf{B}}(t;\boldsymbol{\beta}) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})\mathbf{E}_\mathbf{r}(\boldsymbol{\beta},u)\left\{S_\mathbf{r}^{(0)}(\boldsymbol{\beta},u)\right\}^{-1} \mathrm{d}N_\mathbf{r}(u). \tag{6.24}$$

Then, by (2.13) - (2.15) and (3.4)

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \left\|\widehat{\mathbf{B}}(t;\boldsymbol{\beta}^*) - \mathbf{B}(t;\boldsymbol{\beta}_0)\right\| \\
&\leq \sup_{t\in[0,\tau]} \left\|\widehat{\mathbf{B}}(t;\boldsymbol{\beta}^*) - \widehat{\mathbf{B}}(t;\boldsymbol{\beta}_0)\right\| \\
&\quad + \sup_{t\in[0,\tau]} \left\|\int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})\mathbf{E}_\mathbf{r}(\boldsymbol{\beta}_0,u)\left\{S_\mathbf{r}^{(0)}(\boldsymbol{\beta}_0,u)\right\}^{-1} \mathrm{d}M_\mathbf{r}(u)\right\| \\
&\quad + \sup_{t\in[0,\tau]} \left\|\int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})\mathbf{E}_\mathbf{r}(\boldsymbol{\beta}_0,u)\alpha_0(u)\mathrm{d}u - \int_0^t e(\boldsymbol{\beta}_0,u)\alpha_0(u)\mathrm{d}u\right\|.
\end{aligned} \tag{6.25}$$

We will show that each of the three terms on the right hand side converge to zero in probability,

For the first term we use a Taylor series expansion to get for the j th component $\widehat{B}^j(t;\boldsymbol{\beta})$ of $\widehat{\mathbf{B}}(t;\boldsymbol{\beta})$ when $\boldsymbol{\beta}^* \in \mathcal{B}_0$

$$\widehat{B}^j(t;\boldsymbol{\beta}^*) = \widehat{B}^j(t;\boldsymbol{\beta}_0) + \sum_{k=1}^p (\beta_k^* - \beta_{k0}) \frac{\partial}{\partial\beta_k}\widehat{B}^j(t;\check{\boldsymbol{\beta}}),$$

with $\check{\boldsymbol{\beta}}$ on the line segment joining $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_0$. By (6.7), (6.8) and (6.24)

$$\left|\frac{\partial}{\partial\beta_k}\widehat{B}^j(t;\boldsymbol{\beta})\right| \leq 2\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})X_{\mathbf{r},n}(u)^2\left\{S_\mathbf{r}^{(0)}(\boldsymbol{\beta},u)\right\}^{-1} \mathrm{d}N_\mathbf{r}(u).$$

Therefore (6.5), Condition 7 and an application of Lenglart's inequality similar to (6.17) give that, for all j,k,l, $t\in[0,\tau]$ and $\boldsymbol{\beta}\in\mathcal{B}_0$, $\left|\frac{\partial}{\partial\beta_k}\widehat{B}^j(t;\boldsymbol{\beta})\right|$ is bounded in probability by a finite constant K' not depending on $\boldsymbol{\beta}$. Thus the leading term on the right hand side of (6.25) is bounded in probability by $pK'\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0\|$ and therefore tends to zero in probability as $\boldsymbol{\beta}^* \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_0$.

The predictable variation process, evaluated at $t=\tau$, of the stochastic integral in the second term of (6.25) is bounded by $1/n$ times a constant times (6.20) with $\gamma = 1$. That this term converges to zero in probability therefore follows by Condition 7 and an application of Lenglart's inequality similar to (6.16). Finally, the third term on the right hand side of (6.25) is bounded by

$$\int_0^\tau \left\|\sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})\mathbf{E}_\mathbf{r}(\boldsymbol{\beta}_0,u) - e(\boldsymbol{\beta}_0,u)\right\| \alpha_0(u)\mathrm{d}u,$$

which tends to zero in probability by dominated convergence (Proposition 1) invoking (6.7) and Condition 7. □

Following Andersen and Gill (1982, Theorem 3.4), see also Andersen *et al.* (1992, Theorem VII.2.3), we may then prove the following result about the asymptotic joint distribution of the estimator (4.3) and $\widehat{\boldsymbol{\beta}}$:

Theorem 4 *Assume Conditions 1-8. Then, with* $\phi(\beta_0, t)$ *and* $\mathbf{B}(t;\beta_0)$ *defined* (6.22) *and* (6.23) , *the processes*

$$W(\cdot) = \sqrt{n}\left(\hat{A}_0(\cdot;\hat{\beta}) - A_0(\cdot)\right) + \sqrt{n}\left(\hat{\beta} - \beta_0\right)^{\mathsf{T}} B(\cdot;\beta_0)$$

and $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ *are asymptotically independent, and* W *converges weakly to a mean zero Gaussian martingale with variance function*

$$\omega^2(t) = \int_0^t \phi(\beta_0, u)\alpha_0(u)\mathrm{d}u. \tag{6.26}$$

Proof: By (4.5) we may write

$$\sqrt{n}\left(\hat{A}_0(t;\hat{\beta}) - A_0(t)\right) = \tag{6.27}$$
$$\sqrt{n}\left(\hat{A}_0(t,\hat{\beta}) - \hat{A}_0(t;\beta_0)\right) + \sqrt{n}\widehat{W}(t) + \sqrt{n}\int_0^t (J(u) - 1)\alpha_0(u)\mathrm{d}u,$$

with $\widehat{W}$ defined by (4.6). Here the last term converges (uniformly in t) in probability to zero by Conditions 1 and 6, while the leading term, by a Taylor series expansion, equals

$$-\sqrt{n}\left(\hat{\beta} - \beta_0\right)^{\mathsf{T}} \hat{\mathbf{B}}(t;\beta^*)$$

with β^* between $\hat{\beta}$ and β_0. Invoking Lemma 1 it follows that the leading term on the right hand side of (6.27) is asymptotically equivalent to $\sqrt{n}(\hat{\beta} - \beta_0)^{\mathsf{T}}\mathbf{B}(t;\beta_0)$.

Furthermore the predictable covariation process between the local square integrable martingales $n^{-1/2}\mathbf{U}_\cdot(\beta_0)$ and $\sqrt{n}\widehat{W}$ is (use (2.7)-(2.10), (3.4), (3.10), (4.1) and (4.6))

$$\langle n^{-1/2}\mathbf{U}_\cdot(\beta_0), \sqrt{n}\widehat{W}\rangle(t) =$$
$$\int_0^t \sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}} \{\mathbf{Z}_i(u) - \mathbf{E}_{\mathbf{r}}(\beta_0, u)\}J(u)\pi_u(\mathbf{r})\{S_{\mathbf{r}}^{(0)}(\beta_0, u)\}^{-1}\lambda_{(i,\mathbf{r})}(u)\mathrm{d}u = 0,$$

i.e. they are orthogonal and therefore asymptotically independent. But by (6.18) and Theorems 1 and 2

$$\sqrt{n}(\hat{\beta} - \beta_0) = \mathbf{\Sigma}^{-1}\frac{1}{\sqrt{n}}\mathbf{U}_\tau(\beta_0) + o_p(1),$$

so that $\sqrt{n}(\hat{\beta} - \beta_0)$ and $\sqrt{n}\widehat{W}$ are asymptotically independent as well.

By (6.27) it therefore only remains to prove that $\sqrt{n}\widehat{W}$ converges weakly to a Gaussian martingale with covariance function ω^2 given by (6.26). But this follows by the martingale central limit theorem (e.g. Andersen *et al.*, 1992, Section II.5.1) since by (2.13) - (2.16), (3.4), (4.1) and (4.6)

$$\langle\sqrt{n}\widehat{W}\rangle(t) = n\int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})^2\left\{S_{\mathbf{r}}^{(0)}(\beta_0, u)\right\}^{-1}\alpha_0(u)\mathrm{d}u, \tag{6.28}$$

and this tends in probability to $\omega^2(t)$ for all $t \in [0,\tau]$ by dominated convergence (Proposition 1) invoking (6.5) and Conditions 7 and 8. For the Lindeberg condition a Chebychev type inequality gives for any $\epsilon > 0$

$$n\int_0^\tau \sum_{r\in\mathcal{P}} J(u)\pi_u(r)^2 \left\{S_r^{(0)}(\beta_0,u)\right\}^{-1} I\left\{\sqrt{n}J(u)\pi_u(r)\left(S_r^{(0)}(\beta_0,u)\right)^{-1} > \epsilon\right\}\alpha_0(u)\mathrm{d}u$$
$$\leq (n^{3/2}/\epsilon)\int_0^\tau \sum_{r\in\mathcal{P}} J(u)\pi_u(r)^3\left\{S_r^{(0)}(\beta_0,u)\right\}^{-2}\alpha_0(u)\mathrm{d}u,$$

which converges to zero in probability by (6.5) and Condition 7. □

As a consequence of Theorem 4 we get the following result for the asymptotic distribution of $\widehat{A}(t;\mathbf{Z}_0)$ defined in (4.7) (cf. Andersen *et al.*, 1992, Corollary VII.2.6)

Theorem 5 *Assume Conditions 1-8. Then the process* (4.8) *converges weakly to a mean zero Gaussian process with a covariance function*

$$\sigma^2(s,t) = \left\{\exp(\beta_0^\mathsf{T}\mathbf{Z}_0)\right\}^2 \tag{6.29}$$
$$\times\left\{\omega^2(s\wedge t;\beta_0) + (\mathbf{B}(s;\beta_0) - A_0(s)\mathbf{Z}_0)^\mathsf{T}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{B}(t;\beta_0) - A_0(t)\mathbf{Z}_0)\right\}$$

which may be estimated uniformly consistently by $n\widehat{\sigma}^2(s,t)$, *cf.* (4.9).

Proof: As a consequence of Theorems 3 and 4 the process $\sqrt{n}(\widehat{A}_0(\cdot\,;\widehat{\beta}) - A_0(\cdot))$ converges weakly to a mean zero Gaussian process with a covariance function

$$\omega^2(s\wedge t;\beta_0) + \mathbf{B}(s;\beta_0)^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{B}(t;\beta_0). \tag{6.30}$$

It also follows that the asymptotic covariance of $\sqrt{n}(\widehat{\beta}-\beta_0)^\mathsf{T}$ and $\sqrt{n}(\widehat{A}_0(t\,;\widehat{\beta}) - A_0(t))$ is

$$-\boldsymbol{\Sigma}^{-1}\mathbf{B}(t;\beta_0). \tag{6.31}$$

Now, by a Taylor series expansion, the processes (4.8) and

$$\exp\left(\beta_0^\mathsf{T}\mathbf{Z}_0\right)\left\{\sqrt{n}\left(\widehat{A}_0(\cdot\,;\widehat{\beta}) - A_0(\cdot)\right) + A_0(\cdot)\mathbf{Z}_0^\mathsf{T}\sqrt{n}(\widehat{\beta}-\beta_0)\right\}$$

are asymptotically equivalent, and the weak convergence result for (4.8) follows invoking (6.30) and (6.31).

Finally, we will prove that $n\widehat{\sigma}^2(s,t)$ is a uniformly (in s and t) consistent estimator for (6.29). By Theorems 1, 2 and 4 and Lemma 1, we then only need to prove that $n\widehat{\omega}^2(t;\widehat{\beta})$ (cf. (4.10)) is a uniformly consistent estimator for (6.26). Now $n\widehat{\omega}^2(\cdot\,;\beta_0)$ is the optional variation process of the local square integrable martingale $\sqrt{n}\widehat{W}$ (cf. (4.6)) and therefore, by Rebolledo's theorem (cf. Andersen *et al.*, 1992, Theorem II.5.1), tends uniformly in probability to the same limit as the predictable variation process of this martingale. Thus, by (6.28),

$$\sup_{t\in[0,\tau]} \|n\widehat{\omega}^2(t;\beta_0) - \omega^2(t)\| \xrightarrow{\mathrm{P}} 0$$

as $n \to \infty$. Furthermore, by (6.7),

$$\left|\frac{\partial}{\partial \beta_k}\widehat{\omega}^2(t;\beta)\right| \leq 2\int_0^\tau \sum_{\mathbf{r}\in\mathcal{P}} J(u)\pi_u(\mathbf{r})^2 X_{\mathbf{r},n}(u)\left\{S_{\mathbf{r}}^{(0)}(\beta_0,u)\right\}^{-2} \mathrm{d}N_{\mathbf{r}}(u),$$

and it follows by (6.5) and Condition 7, using an argument similar to the one used to handle the first term on the right hand side of (6.25), that

$$\sup_{t\in[0,\tau]} \|n\omega^2(t;\widehat{\beta}) - n\omega^2(t;\beta_0)\| \xrightarrow{\mathrm{P}} 0,$$

and the uniform consistency of $n\omega^2(t;\widehat{\beta})$ follows. □

7 Asymptotic results for specific sampling schemes

In this section, we will assume throughout Conditions 1 and 2, and we will demonstrate how Conditions 3 - 8 are satisfied for some specific sampling schemes under the assumption that the censoring indicators and covariate processes $(Y_i(\cdot), \mathbf{Z}_i(\cdot)), i \in \{1,2,\ldots,n\}$ are independent copies of the pair $(Y, \mathbf{Z})$. In all cases below we let $p(t) = P(Y(t) = 1)$, and we assume $\inf_{t\in[0,\tau]} p(t) > 0$ and $\alpha_0(t) > 0$ for almost all $t \in [0,\tau]$. We also assume that

$$\mathbf{V} = \int_0^\tau \mathrm{Cov}(\mathbf{Z}_Y(t))\alpha_0(t)\mathrm{d}t \quad \text{is positive definite,} \tag{7.1}$$

where the distribution of $\mathbf{Z}_Y$ is given by

$$P(\mathbf{Z}_Y(t) \in B) = P(\mathbf{Z}(t) \in B | Y(t) = 1).$$

Our Conditions 3 - 8 are taylored for cohort sampling methods, and they are more restrictive than necesary for the situation of Example 5.1 where the entire risk set is sampled with probability one. Nevertheless it may be illustrative to first consider this simple example.

Example 7.1 Full cohort. In order to have Conditions 5 and 7 fulfilled we here need to assume that the covariate process $\mathbf{Z}$ is bounded; that is, that there exists M such that $\|\mathbf{Z}(t)\| \leq M$ for all $t \in [0,\tau]$.

Under the above assumptions we will show how Conditions 3, 4 and 5 are satisfied; hence Theorems 1,2 and 3 may be invoked. In full cohort sampling the entire risk set is used as controls for the case so that $\pi_t(\mathbf{r} \mid i) = I(\mathbf{r} = \mathcal{R}(t))$ for all $i \in \mathcal{R}(t) = \{i : Y_i(t) = 1\}$. In this case (3.4) and (3.5) become the cohort quantities

$$S^{(\gamma)}(\beta,t) = \sum_{i=1}^n Y_i(t)\mathbf{Z}_i^{\otimes\gamma}(t)\exp(\beta^\mathsf{T}\mathbf{Z}_i(t))$$

and

$$\mathbf{E}(\beta,t) = S^{(1)}(\beta,t)/S^{(0)}(\beta,t),$$

respectively, cf. Andersen and Gill (1982). By (6.1),

$$Q^{(\rho,\gamma)}(\beta_0,t)=\left(\frac{n^{-1}S^{(1)}(\beta_0,t)}{n^{-1}S^{(0)}(\beta_0,t)}\right)^{\otimes\rho}\frac{1}{n}S^{(\gamma)}(\beta_0,t).$$

The law of large numbers yields

$$\frac{1}{n}S^{(\gamma)}(\beta,t)\xrightarrow{\mathrm{P}}s^{(\gamma)}(\beta,t)\quad\text{as } n\to\infty, \tag{7.2}$$

where

$$s^{(\gamma)}(\beta_0,t)=E\{Y(t)\mathbf{Z}^{\otimes\gamma}(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}(t))\}.$$

Taking limits,

$$q^{(\rho,\gamma)}(\beta_0,t)=\left(\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)^{\otimes\rho}s^{(\gamma)}(\beta_0,t).$$

In particular,

$$q^{(0,2)}(\beta_0,t)=s^{(2)}(\beta_0,t)\quad\text{and}\quad q^{(2,0)}(\beta_0,t)=s^{(1)}(\beta_0,t)^{\otimes 2}/s^{(0)}(\beta_0,t),$$

and

$$\mathbf{\Sigma}=\int_0^\tau[s^{(2)}(\beta_0,t)/s^{(0)}(\beta_0,t)-\{s^{(1)}(\beta_0,t)/s^{(0)}(\beta_0,t)\}^{\otimes 2}]s^{(0)}(\beta_0,t)\alpha_0(t)\mathrm{d}t,$$

recovering the covariance given by Andersen and Gill (1982).

As the covariate processes are bounded, (6.3) is satisfied with $X_{\mathbf{r}}(t)=M$. Noting that in the full cohort case $\sum_{\mathbf{r}\in\mathcal{P}}\sum_{i\in\mathbf{r}}\pi_t(\mathbf{r}|i)=|\mathcal{R}(t)|=n(t)$, we see that (6.4) is satisfied as $n(t)/n\xrightarrow{\mathrm{P}}p(t)$ and $\int_0^\tau(n(t)/n)\alpha_0(t)\mathrm{d}t$ converges in probability to $\int_0^\tau p(t)\alpha_0(t)\mathrm{d}t$, which is finite by Condition 1. Thus Condition 5 is fulfilled.

We observe that the covariance $\mathbf{\Sigma}$ has the following interpretation. Define the distribution of a random vector $\widehat{\mathbf{Z}}_Y(t)$ by

$$Eh(\widehat{\mathbf{Z}}_Y(t))=E\left(h(\mathbf{Z}_Y(t))\frac{\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_Y(t))}{E\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_Y(t))}\right), \tag{7.3}$$

for all bounded measurable functions h. Then

$$\mathbf{\Sigma}=\int_0^\tau p(t)\mathrm{Cov}(\widehat{\mathbf{Z}}_Y(t))E\{\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_Y(t))\}\alpha_0(t)\mathrm{d}t, \tag{7.4}$$

that is, $\mathbf{\Sigma}$ is the integral of the product of the at risk probability, the covariance of the covariate of an observed failure, and the average hazard for an individual in the cohort.

We now show how condition (7.1) implies Condition 4. If $\mathbf{\Sigma}$ given by (7.4) is not positive definite, then there exists a nonzero vector $\mathbf{a}$ such that $\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{a}=0$, and so by the positivity assumptions, $\mathbf{a}^{\mathsf{T}}\mathrm{Cov}(\widehat{\mathbf{Z}}_Y(t))\mathbf{a}=0$ for almost all t. Hence, for almost all t, $E([\mathbf{a}^{\mathsf{T}}(\widehat{\mathbf{Z}}_Y(t)-E\widehat{\mathbf{Z}}_Y(t))]^2=0$, which, by (7.3), implies that for almost all t, $E([\mathbf{a}^{\mathsf{T}}(\mathbf{Z}_Y(t)-E\mathbf{Z}_Y(t))]^2\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_Y(t))/E(\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_Y(t)))=0$. But then, for such t, $E[\mathbf{a}^{\mathsf{T}}(\mathbf{Z}_Y(t)-E\mathbf{Z}_Y(t))]^2=0$, and therefore $\mathbf{a}^{\mathsf{T}}\mathrm{Cov}(\mathbf{Z}_Y(t))\mathbf{a}=0$ and $\mathbf{V}$ is not positive definite, contrary to assumption (7.1).

To satisfy the hypotheses of Theorems 4 and 5, we find it necessary to impose a further condition. An example given in Goldstein and Langholz (1992) demonstrates that $\inf_{t\in[0,\tau]}p(t)>0$ is not sufficient to imply Condition 6. Therefore, we assume

Condition 9 *There exists a finite number of intervals* $I_k, k = 1, \ldots, K$, *such that*

$$[0, \tau] \subset \cup_{k=1}^{K} I_k$$

and with $p_k = P(Y(t) = 1 \forall t \in I_k)$, *we have* $\min_k p_k > 0$.

Then Condition 6 is fulfilled, for with $J_k = I(\sum_{i=1}^{n} Y_i(t) > 0 \forall t \in I_k)$ we have that $J(t) \geq \prod_{k=1}^{K} J_k$. But $P(J_k = 1) = 1 - (1 - p_k)^n$; hence J_k and therefore $J(t)$ converge to 1 in probability.

Moreover Condition 7 is seen to be satisfied by an argument similar to the one used to prove Condition 5, while (7.2) yields that the left hand sides of (6.21) and (6.22) converge in probability to

$$e(\beta_0, t) = s^{(1)}(\beta_0, t)/s^{(0)}(\beta_0, t)$$

and

$$\phi(\beta_0, t) = 1/s^{(0)}(\beta_0, t),$$

respectively, recovering the results of Andersen and Gill (1982, Theorem 3.4).

We next consider nested case-control sampling where, at each failure, $m - 1$ controls are randomly sampled from the risk set (Example 5.2).

Example 7.2 Nested case-control sampling. We here assume that the covariate process **Z** satisfies the following moment condition:

Condition 10 *There exists* $b_0 > 3||\beta_0||$ *such that,*

$$E\{\exp(2b_0||\mathbf{Z}(t)||)\} < \infty$$

for all $t \in [0, \tau]$, *and*

$$\int_0^\tau \{E \exp(2b_0||\mathbf{Z}(t)||)\}^{m/2} \alpha_0(t) \mathrm{d}t < \infty.$$

Using the fact that if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $E\{\exp(\gamma|Z|)\} \leq 2\exp(\frac{1}{2}\gamma^2\sigma^2 + \gamma|\mu|)$, it is not difficult to see that Condition 10 is satisfied if, for example,

$$Z(t) \sim \mathcal{N}(\mu(t), \sigma^2(t))$$

with

$$\sup_{t \in [0,\tau]} |\mu(t)| < \infty \quad \text{and} \quad \sup_{t \in [0,\tau]} \sigma^2(t) < \infty$$

under Condition 1.

We will show that Condition 10 and the above assumptions imply that Conditions 3 through 5 are satisfied, and so the conclusions of Theorems 1, 2 and 3.

Let $\mathcal{R}(t)$ and $n(t)$ as above be the risk set and the number at risk at time t, and introduce

$$\mathcal{P}(t) = \{\mathbf{r} \subset \mathcal{R}(t) : |\mathbf{r}| = m\}, \quad \mathcal{P}_i(t) = \{\mathbf{r} \in \mathcal{P}(t), i \in \mathbf{r}\},$$

and

$$\pi_t(\mathbf{r}|i) = \binom{n(t)-1}{m-1}^{-1} I(\mathbf{r} \in \mathcal{P}_i(t)).$$

First we verify Condition 3, which gives us the form of the asymptotic covariance matrix. For (ρ, γ) equal either (0,2) or (2,0), we have by (6.1)

$$Q^{(\rho,\gamma)}(\boldsymbol{\beta}_0, t) = \frac{1}{n}\binom{n(t)-1}{m-1}^{-1} \sum_{\mathbf{r}\in\mathcal{P}} \left\{ \mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}_0, t)^{\otimes\rho} \left[\sum_{i\in\mathbf{r}} \mathbf{Z}_i(t)^{\otimes\gamma} \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_i(t)) \right] I(\mathbf{r} \in \mathcal{P}(t)) \right\}.$$

Let

$$\overline{Q}^{(\rho,\gamma)}(\boldsymbol{\beta}_0, t) = p(t)^{-m+1} \frac{1}{m}\binom{n}{m}^{-1} \sum_{\mathbf{r}\in\mathcal{P}} \left\{ \mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}_0, t)^{\otimes\rho} \left[\sum_{i\in\mathbf{r}} \mathbf{Z}_i(t)^{\otimes\gamma} \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_i(t)) \right] I(\mathbf{r} \in \mathcal{P}(t)) \right\}.$$

In calculating the variance of $\overline{Q}$, we have $\binom{n}{m}^2 - \binom{n}{2m}\binom{2m}{m}$ nonzero terms corresponding to the number of sets $\mathbf{r}, \mathbf{s}$ where $|\mathbf{r}| = |\mathbf{s}| = m$ and $\mathbf{r} \cap \mathbf{s} \neq \emptyset$. Since $\lim_{n\to\infty} \binom{n}{m}^{-2}\binom{n}{2m}\binom{2m}{m} = 1$, we see that $\mathrm{Var}\overline{Q} \to 0$, and therefore $\overline{Q}$ converges to its expectation (Goldstein and Langholz, 1992) . Taking ratios componentwise if necessary and using $n(t)/n \xrightarrow{\mathrm{P}} p(t)$ as $n \to \infty$, we see that $\overline{Q}$ and Q have the same limit in probability. Hence with $\mathbf{u} = \{1, 2, \ldots, m\}$,

$$q^{(\rho,\gamma)}(\boldsymbol{\beta}_0, t) = p(t)^{-m+1} E\left\{ \mathbf{E}_{\mathbf{u}}(\boldsymbol{\beta}_0, t)^{\otimes\rho} \left[\frac{1}{m} \sum_{i\in\mathbf{u}} \mathbf{Z}_i(t)^{\otimes\gamma} \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_i(t)) \right] I(\mathbf{u} \in \mathcal{P}(t)) \right\}.$$

In particular,

$$\boldsymbol{\Sigma} = \int_0^1 [q^{(0,2)}(\boldsymbol{\beta}_0, t) - q^{(2,0)}(\boldsymbol{\beta}_0, t)]\alpha_0(t)dt,$$

where

$$\begin{aligned} & q^{(0,2)}(\boldsymbol{\beta}_0, t) - q^{(2,0)}(\boldsymbol{\beta}_0, t) \\ &= p(t)^{-m+1} E\left\{ \left(\mathbf{Z}(t)^{\otimes 2} \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}(t)) - \frac{1}{m} \mathbf{E}_{\mathbf{u}}(\boldsymbol{\beta}_0, t)^{\otimes 2} \right) I(\mathbf{u} \in \mathcal{P}(t)) \right\}. \end{aligned}$$

As shown in Goldstein and Langholz (1992), Condition 4 is satisfied whenever condition (7.1) is true by an argument similar to that used in the full cohort case.

To interpret the covariance matrix for this instance, let $\mathbf{Z}_{Y,\mathbf{u}} = (\mathbf{Z}_{Y,1}, \mathbf{Z}_{Y,2}, \ldots, \mathbf{Z}_{Y,m})^{\mathsf{T}}$, a vector with independent components each with distribution $\mathbf{Z}_Y$. Let the distribution of $\widehat{\mathbf{Z}}_Y$ be specified by

$$P(\widehat{\mathbf{Z}}_Y(t) = \mathbf{Z}_{Y,j}(t) | \mathbf{Z}_{Y,\mathbf{u}}) = q_j(t)$$

where

$$q_j(t) = \frac{\exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_{Y,j}(t))}{\sum_{i\in\mathbf{u}} \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_{Y,i}(t))}.$$

Then

$$\Sigma = E\left\{\int_0^\tau p(t)\mathrm{Cov}(\hat{\mathbf{Z}}_Y(t)|\mathbf{Z}_{Y,\mathbf{u}})\frac{1}{m}\sum_{i\in\mathbf{u}}\exp(\beta_0^\mathsf{T}\mathbf{Z}_{Y,i}(t))\alpha_0(t)\mathrm{d}t\right\}, \tag{7.5}$$

which has a similar interpretation to that given to (7.4) in the case of the full cohort; the matrix Σ is the expectation of the integral of the product of the at risk probability, an estimate of the covariance of the covariate of the failure, and an estimate of the average hazard of the cohort, both based on the sampled risk set.

Next we verify Condition 5. Equation (6.3) is true using the bounding random variables

$$X_{\mathbf{r}}(t) = \sum_{i\in\mathbf{r}}||\mathbf{Z}_i(t)||.$$

To verify (6.4), begin by noting that

$$D_n(t) = \frac{n(t)}{n}\binom{n(t)}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\left\{e^{b_0 X_{\mathbf{r}}(t)}I(\mathbf{u}\in\mathcal{P}(t))\right\}.$$

The set counting argument used to show Condition 3 may be repeated to show that $D_n(t) \xrightarrow{\mathrm{P}} D(t)$ where $D(t) = p(t)^{-m+1}E\{e^{b_0 X_{\mathbf{u}}(t)}I(\mathbf{u}\in\mathcal{P}(t))\}$. The verification of (6.4) will therefore be complete if we show

$$\begin{aligned}&\int_0^\tau \frac{n(t)}{n}\binom{n(t)}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}} e^{b_0 X_{\mathbf{r}}(t)}I(\mathbf{r}\in\mathcal{P}(t))\alpha_0(t)\mathrm{d}t\\ &\xrightarrow{\mathrm{P}} \int_0^\tau p(t)^{-m+1}E\{e^{b_0 X_{\mathbf{u}}(t)}I(\mathbf{u}\in\mathcal{P}(t))\}\alpha_0(t)\mathrm{d}t,\end{aligned}$$

this latter quantity being finite by $\inf_t p(t) > 0$ and Condition 10. Let $\epsilon > 0$ be arbitrary and defining the events

$$B_t = \left\{\left|\frac{n(t)}{n} - p(t)\right| \le \epsilon\right\},$$

there exists a constant $C > 0$, depending only on ϵ, such that

$$P(B_t^c) < Ce^{-n/C}. \tag{7.6}$$

Recall that if $n(t) < m$ the probability of sampling a set $\mathbf{r}$ is zero; we may therefore adopt the convention that $\binom{n(t)}{m}^{-1} = 0$ whenever $n(t) < m$. We now see that

$$\int_0^\tau \frac{n(t)}{n}\binom{n(t)}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}} e^{b_0 X_{\mathbf{r}}(t)}I(B_t^c)I(\mathbf{r}\in\mathcal{P}(t))\alpha_0(t)\mathrm{d}t \xrightarrow{\mathrm{P}} 0$$

using $|\frac{n(t)}{n}\binom{n(t)}{m}^{-1}| \le 1$, taking expectation, and applying the Cauchy-Schwarz inequality, Condition 10 and (7.6). Since

$$\frac{N}{n}\binom{N}{m}^{-1}\binom{n}{m} = \binom{N}{n}^{-m+1} + O(1/n),$$

we have

$$\frac{n(t)}{n}\binom{n(t)}{m}^{-1}\binom{n}{m}I(B_t) = p(t)^{-m+1}I(B_t) + O(\epsilon + 1/n),$$

and the result will follow from

$$\int_0^\tau p(t)^{-m+1}\binom{n}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}} e^{b_0 X_{\mathbf{r}}(t)} I(\mathbf{r}\in\mathcal{P}(t))\alpha_0(t)\mathrm{d}t$$
$$\xrightarrow{\mathrm{P}} \int_0^\tau p(t)^{-m+1}E\{e^{b_0 X_{\mathbf{u}}(t)}I(\mathbf{u}\in\mathcal{P}(t))\}\alpha_0(t)\mathrm{d}t.$$

Computing the second moment of the difference and again using $\inf_t p(t) > 0$, it suffices to show

$$\int_0^\tau\int_0^\tau \binom{n}{m}^{-2}\sum_{\substack{|\mathbf{r}|=m\\|\mathbf{s}|=m}} \mathrm{Cov}\{e^{b_0 X_{\mathbf{r}}(t)}I(\mathbf{r}\in\mathcal{P}(t)), e^{b_0 X_{\mathbf{s}}(u)}I(\mathbf{s}\in\mathcal{P}(u))\}\alpha_0(t)\alpha_0(u)\mathrm{d}t\mathrm{d}u$$

tends to zero as $n\to\infty$. Again, the only nonzero terms in the double sum above are those for which $\mathbf{r}\cap\mathbf{s}\neq\emptyset$, and the same set counting argument as above shows that it suffices to prove

$$\int_0^\tau\int_0^\tau \mathrm{Cov}\{e^{b_0 X_{\mathbf{r}}(t)}I(\mathbf{r}\in\mathcal{P}(t)), e^{b_0 X_{\mathbf{s}}(u)}I(\mathbf{s}\in\mathcal{P}(u))\}\alpha_0(t)\alpha_0(u)\mathrm{d}t\mathrm{d}u < \infty,$$

which follows from Condition 10 after a final application of the Cauchy-Schwarz inequality.

To satisfy the hypotheses of Theorems 4 and 5, we assume Condition 9 of the previous example to hold. Then Condition 6 is fulfilled. To verify Condition 7, we see that in the present case equation (6.20) reduces to

$$\int_0^\tau J(t)\left(\frac{n(t)}{n}\right)^{-\gamma}\binom{n(t)}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\{\exp(b_0 X_{\mathbf{r}}(t))I(\mathbf{r}\in\mathcal{P}(t))\}\,\alpha_0(t)\mathrm{d}t.$$

The calculation that this quantity converges in probability to a finite quantity parallels that for showing the convergence in probability of $\int_0^\tau D_n(t)\alpha_0(t)\mathrm{d}t$.

For Condition 8, using $J(t)\xrightarrow{\mathrm{P}} 1$ and arguments similar to those used above we have equation (6.21) satisfied with

$$e(\beta_0,t) = p(t)^{-m}E\left\{\mathbf{E}_{\mathbf{u}}(\beta_0,t)I(\mathbf{u}\in\mathcal{P}(t))\right\}$$

Similarly, since the left hand side of (6.22) reduces to

$$\frac{J(t)mn}{n(t)}\binom{n(t)}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\left\{\left(\sum_{i\in\mathbf{r}}Y_i(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t))\right)^{-1}I(\mathbf{r}\in\mathcal{P}(t))\right\},$$

we see equation (6.22) is satisfied with

$$\phi(\beta_0,t) = \frac{m}{p(t)^{m+1}}E\left\{\left(\sum_{i\in\mathbf{u}}Y_i(t)\exp(\beta_0^{\mathsf{T}}\mathbf{Z}_i(t))\right)^{-1}I(\mathbf{u}\in\mathcal{P}(t))\right\}.$$

From these expressions the asymptotic covariance function (6.29) of the estimated integrated hazard may be obtained using (6.23) and (6.26).

We finally consider stratified nested case-control sampling as discussed in Example 5.3.

Example 7.3 Stratified nested case-control sampling. Assume here that $(Y_i(\cdot), \mathbf{Z}_i(\cdot), C_i(\cdot))$, $i = 1, 2, \ldots, n$ are independent copies of $(Y(\cdot), \mathbf{Z}(\cdot), C(\cdot))$ with $C(t) \in \mathcal{C} = \{1, \ldots, L\}$ and that the covariate process satisfy Condition 10 of the previous example. Note that the independent C assumption covers cases where stratification is done based on observed information on the ith individual only, such as stratification based on absolute exposure status as discussed in Example 5.3, yet, it does not cover stratification based on relative exposure, as the strata of the ith individual now depends on values observed on other cohort members.

Using the notation of Example 5.3, we let

$$p_l(t) = P(C(t) = l | Y(t) = 1) \quad l = 1, 2, \ldots, L,$$

$$\mathcal{P}(t) = \{\mathbf{r} \subset \mathcal{R}(t) : \forall\, 0 \leq l \leq L, |\mathbf{r} \cap \mathcal{R}_l(t)| = m_l\},$$

$$\mathcal{P}_i(t) = \{\mathbf{r} \in \mathcal{P}(t), i \in \mathbf{r}\},$$

and

$$\pi_t(\mathbf{r}|i) = \left[\prod_{l=1}^{L} \binom{n_l(t)}{m_l}^{-1}\right] \frac{n_{C_i(t)}(t)}{m_{C_i(t)}} I(\mathbf{r} \in \mathcal{P}_i(t)).$$

We assume that $\inf_{t \in [0,\tau]} p_l(t) > 0, l = 1, 2, \ldots L$. Note that $\mathcal{P}(t)$ and $\mathcal{P}_i(t)$ have different meanings than in the previous example.

We first consider Condition 3. For (ρ, γ) equal either (2,0) or (0,2), we have by (6.1)

$$Q^{(\rho,\gamma)}(\beta_0, t) = \frac{1}{n} \sum_{\mathbf{r} \in \mathcal{P}} \mathbf{E}_{\mathbf{r}}(\beta_0, t)^{\otimes \rho} \sum_{i \in \mathbf{r}} \mathbf{Z}_i(t)^{\otimes \gamma} \exp(\beta_0^{\mathsf{T}} \mathbf{Z}_i(t)) \pi_t(\mathbf{r}|i).$$

Let

$$f(t) = p(t)^{-m+1} \binom{m}{m_1, \ldots, m_L}^{-1} \prod_{l=1}^{L} p_l(t)^{-m_l},$$

and

$$r_t(\mathbf{r}|i) = f(t) \frac{p_{C_i(t)}(t)}{m_{C_i(t)}} \binom{n}{m}^{-1} I(\mathbf{r} \in \mathcal{P}_i(t)).$$

Since $n_l(t)/n \xrightarrow{\mathrm{P}} p(t)p_l(t)$ as $n \to \infty$,

$$\frac{\pi_t(\mathbf{r}|i)}{n r_t(\mathbf{r}|i)} \xrightarrow{\mathrm{P}} 1 \quad \text{as } n \to \infty,$$

and an argument as in Example 7.2 above now yields that the desired limit is the same as the limit of

$$\overline{Q}^{(\rho,\gamma)}(\boldsymbol{\beta}_0,t)$$

$$= f(t)\binom{n}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\left\{\mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}_0,t)^{\otimes\rho}\left[\sum_{i\in\mathbf{r}}\mathbf{Z}_i(t)^{\otimes\gamma}\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_i(t))p_{C_i(t)}(t)m_{C_i(t)}^{-1}\right]I(\mathbf{r}\in\mathcal{P}(t))\right\}.$$

Let

$$\tilde{S}_{\mathbf{r}}^{(\gamma)}(\boldsymbol{\beta}_0,t)=\sum_{i\in\mathbf{r}}Y_i(t)\mathbf{Z}_i(t)^{\otimes\gamma}\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_i(t))p_{C_i(t)}(t)m_{C_i(t)}^{-1},$$

$$\tilde{\mathbf{E}}_{\mathbf{r}}(\boldsymbol{\beta}_0,t)=\tilde{S}_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}_0,t)/\tilde{S}_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}_0,t)$$

and

$$\tilde{Q}^{(\rho,\gamma)}(\boldsymbol{\beta}_0,t)$$

$$= f(t)\binom{n}{m}^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\left\{\tilde{\mathbf{E}}_{\mathbf{r}}(\boldsymbol{\beta}_0,t)^{\otimes\rho}\left[\sum_{i\in\mathbf{r}}\mathbf{Z}_i(t)^{\otimes\gamma}\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_i(t))p_{C_i(t)}(t)m_{C_i(t)}^{-1}\right]I(\mathbf{r}\in\mathcal{P}(t))\right\}.$$

As $n_l(t)/n(t)$ converges to $p_l(t), l=1,2,\ldots,L$, we have that $\tilde{\mathbf{E}}_{\mathbf{r}}-\mathbf{E}_{\mathbf{r}}\xrightarrow{\mathrm{P}}0$ and $\overline{Q}-\tilde{Q}\xrightarrow{\mathrm{P}}0$.

As in Example 7.2, the variance of the $\tilde{Q}$ tends to zero (see Langholz and Goldstein, 1992), and $\tilde{Q}$ converges to its expectation. Hence,

$$q^{(\rho,\gamma)}(\boldsymbol{\beta}_0,t)$$

$$= f(t)E\left\{\tilde{\mathbf{E}}_{\mathbf{u}}(\boldsymbol{\beta}_0,t)^{\otimes\rho}\left[\sum_{j\in\mathbf{u}}\mathbf{Z}_j(t)^{\otimes\gamma}\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_j(t))p_{C_j(t)}(t)m_{C_j(t)}^{-1}\right]I(\mathbf{u}\in\mathcal{P}(t))\right\},$$

and we have that $\boldsymbol{\Sigma}=\int_0^\tau[q^{(0,2)}(\boldsymbol{\beta}_0,t)-q^{(2,0)}(\boldsymbol{\beta}_0,t)]\alpha_0(t)\mathrm{d}t$.

As in the previous two examples, the condition (7.1) implies Condition 4. The matrix $\boldsymbol{\Sigma}$ here can be interpreted as in the previous two examples, but in this instance, with $\hat{\mathbf{Z}}_Y$ defined below, $\boldsymbol{\Sigma}$ is the expectation of the integral of the product of the at risk probability, an estimate of the covariance of the failure covariate based on the sampled risk set, and an estimate of the average hazard of the entire risk set constructed from the stratified sampled risk set.

To give an expression for $\boldsymbol{\Sigma}$, let (with a slight abuse of notation) $\mathbf{u}=\{(l,j):l=1,2,\ldots,L,j=1,2,\ldots,m_l\}$, $\mathbf{Z}_{Y,\mathbf{u}}(t)=(\mathbf{Z}_{Y,l,j}(t))_{(l,j)\in\mathbf{u}}$ be mutually independent with $\mathbf{Z}_{Y,l,j}(t)$ distributed as $\mathbf{Z}(t)$ conditioned on $Y(t)=1, C(t)=l$, for $(l,j)\in\mathbf{u}$. Now let

$$P(\hat{\mathbf{Z}}_Y(t)=\mathbf{Z}_{Y,l,j}(t)|\mathbf{Z}_{Y,\mathbf{u}})=q_{l,j}(t)$$

where

$$q_{l,j}(t)=\frac{\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{Y,l,j}(t))p_{C_j(t)}m_{C_j(t)}^{-1}}{\sum_{(k,i)\in\mathbf{u}}\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{Y,k,i}(t))p_{C_i(t)}m_{C_i(t)}^{-1}}.$$

Then

$$\Sigma = E\left\{\int_0^\tau p(t)\mathrm{Cov}(\widehat{\mathbf{Z}}_Y(t)|\mathbf{Z}_{Y,\mathbf{u}})\sum_{(k,i)\in\mathbf{u}}\exp(\beta_0^\mathsf{T}\mathbf{Z}_{Y,k,i}(t))p_{C_i(t)}m_{C_i(t)}^{-1}\alpha_0(t)\mathrm{d}t\right\}. \quad (7.7)$$

In particular, formula (7.5) for Σ in nested case-control sampling contains the term $\frac{1}{m}\sum_{i\in\mathbf{u}}\exp(\beta_0^\mathsf{T}\mathbf{Z}_{Y,i}(t))\alpha_0(t)$, which is an estimate of the average hazard in the entire risk set based on the sampled set $\mathbf{u}$. The formula for Σ for stratified nested case-control sampling has the corresponding term

$$\sum_{(k,i)\in\mathbf{u}}\exp(\beta_0^\mathsf{T}\mathbf{Z}_{Y,k,i}(t))p_{C_i(t)}m_{C_i(t)}^{-1}\alpha_0(t).$$

There are $m_{C_i(t)}$ individuals of the same strata as individual i in $\mathbf{u}$, and the factor $m_{C_i(t)}^{-1}$ scales the sum of these contributions to the hazard from these individuals to yield a net hazard for a single individual of this strata. Multiplication of the factor $p_{C_i(t)}$, the proportion of individuals of this strata in the entire risk set, and summing now gives an estimate of the average hazard for the entire risk set based on the stratified sampled risk set. The remainder of terms in the formula for Σ in this case can be similarly interpreted.

To satisfy condition (6.3) again use the bounding random variable

$$X_\mathbf{r}(t) = \sum_{i\in\mathbf{r}}||\mathbf{Z}_i(t)||.$$

The argument that

$$\frac{1}{n}\sum_{\mathbf{r}\in\mathcal{P}}e^{b_0X_\mathbf{r}(t)}\sum_{i\in\mathbf{r}}\pi_t(\mathbf{r}|i)\xrightarrow{\mathrm{P}}f(t)E\left\{e^{b_0X_\mathbf{u}(t)}\left[\sum_{i\in\mathbf{u}}\frac{p_{C_i(t)}(t)}{m_{C_i(t)}}\right]I(\mathbf{u}\in\mathcal{P}(t))\right\}$$

and

$$\begin{aligned}&\int_0^\tau\frac{1}{n}\sum_{\mathbf{r}\in\mathcal{P}}e^{b_0X_\mathbf{r}(t)}\sum_{i\in\mathbf{r}}\pi_t(\mathbf{r}|i)\alpha_0(t)\mathrm{d}t\\&\xrightarrow{\mathrm{P}}\int_0^\tau f(t)E\left\{e^{b_0X_\mathbf{u}(t)}\left[\sum_{i\in\mathbf{u}}\frac{p_{C_i(t)}(t)}{m_{C_i(t)}}\right]I(\mathbf{u}\in\mathcal{P}(t))\right\}\alpha_0(t)\mathrm{d}t\end{aligned}$$

parallels that given in Example 7.2, and verifies condition (6.4).

To satisfy the hypotheses of Theorems 4 and 5, we again assume Condition 9 of the previous example to hold; this implies Condition 6 as before. To verify Condition 7, we see that in the present case equation (6.20) reduces to

$$\int_0^\tau J(t)\left(\frac{n(t)}{n}\right)^{-\gamma}\left[\prod_{l=1}^L\binom{n_l(t)}{m_l}\right]^{-1}\sum_{\mathbf{r}\in\mathcal{P}}\{\exp(b_0X_\mathbf{r}(t))I(\mathbf{r}\in\mathcal{P}(t))\}\alpha_0(t)\mathrm{d}t.$$

which converges in probability to a finite quantity by an argument similar to those used above.

For Condition 8, using $J(t) \xrightarrow{P} 1$ and arguing as before, we have equations (6.21) and (6.22) satisfied with

$$e(\beta_0, t) = p(t)^{-1} f(t) E\left\{\tilde{\mathbf{E}}_{\mathbf{u}}(\beta_0, t) I(\mathbf{u} \in \mathcal{P}(t))\right\}$$

and

$$\phi(\beta_0, t) = f(t) E\left\{\left(\sum_{i \in \mathbf{u}} Y_i(t) \exp(\beta_0^T \mathbf{Z}_i(t)) \frac{p_{C_i(t)}}{m_{C_i(t)}}\right)^{-1} I(\mathbf{u} \in \mathcal{P}(t))\right\}$$

respectively.

Note that Example 7.2 is a special case of Example 7.3 for $L = 1$.

8 Extension to k types of events: stratified populations and multistate models

We have above, for the ease of presentation, assumed that we are interested in only one type of event (failure), and that there is only a single population stratum under consideration. However, in multistate event history analysis, one may want to simultaneously model more than one possible sort of event (e.g. relapse, recovery, death), and in survival analysis (and event history analysis as well) one may want to perform a stratified analysis using some categorical covariate (e.g., sex) as stratification variable. As developed by Andersen and Borgan (1985, Section 7), and further discussed in Andersen *et al.* (1992, Section VII.1), the inclusion of (possible time dependent) strata in our model and the extension to multistate models are both special cases of the same general framework.

For nested case-control sampling this framework may be described as follows. As in Section 2, we first assume that all events observed to happen in the cohort may be modeled as one "large" marked point process. From this marked point process we extract a marked point process $\{(t_j, (h_j, i_j)); j \geq 1\}$ only recording the innovative events, i.e. the times t_j when an event of interest occur, the *type* h_j of the event happening at t_j, and the individual i_j experiencing this event. Here h_j may indicate individual i_j's population stratum, or it may describe what sort of event (e.g. relapse, recovery, death) this individual experiences or both. We denote the possible different types by $1, 2, \ldots, k$, and assume k to be a fixed number not depending on n.

Let

$$N_{hi}(t) = \sum_{j \geq 1} I(t_j \leq t, (h_j, i_j) = (h, i)) \tag{8.1}$$

be the counting process counting the number of events of type h for individual i in $[0, t]$, and let $Y_{hi}(t)$ be a predictable indicator process taking the value 1 if the ith individual is at risk at for experiencing an event of type h just before time t and 0 otherwise. We may then specify a model for the cohort by assuming that the intensity processes of the N_{hi} take the form

$$\lambda_{hi}(t) = Y_{hi}(t) \alpha_{h0}(t) \exp(\beta_0^T \mathbf{Z}_{hi}(t)), \tag{8.2}$$

where the $\mathbf{Z}_{hi}(t)$, for $h = 1, 2, \ldots, k$, are vectors of left continuous and adapted *type specific* covariate processes for the ith individual. These are typically derived from a vector of $\mathbf{Z}_i(t)$ of basic covariates for individual i as illustrated by Andersen *et al.* (1992, Sections VII.1-2).

Sampling may be superimposed onto this cohort model, just as described in Section 2, to get the marked point process

$$\{(t_j, (h_j, i_j, \tilde{\mathcal{R}}_j)); j \geq 1\}$$

also recording the sampled risk sets $\tilde{\mathcal{R}}_j$. This process has mark space

$$E = \{(h, i, \mathbf{r}) : h \in \{1, 2, \ldots, k\}, i \in \{1, 2, \ldots, n\}, \mathbf{r} \in \mathcal{P}_i\}.$$

Also here we let $(\mathcal{F}_t)$ be the filtration generated by the events in the cohort as well as by the sampling. Furthermore define, for each $(h, i, \mathbf{r}) \in E$, the counting process

$$N_{(h,i,\mathbf{r})}(t) = \sum_{j \geq 1} I(t_j \leq t, (h_j, i_j, \mathcal{R}_j) = (h, i, \mathbf{r})), \tag{8.3}$$

counting the observed number of events of type h for individual i in $[0, t]$ with associated sampled risk set equal to $\mathbf{r}$, and denote its intensity process by $\lambda_{(h,i,\mathbf{r})}$.

Assuming the sampling to be independent, we may then write

$$\lambda_{(h,i,\mathbf{r})}(t) = \lambda_{hi}(t)\pi_t(\mathbf{r} | h, i), \tag{8.4}$$

with $\lambda_{hi}(t)$ given by (8.2) and $\pi_t(\mathbf{r} | h, i)$ being the conditional probability of selecting the sampled risk set $\mathbf{r}$ at t given $\mathcal{F}_{t-}$ and the fact that the individual i experiences an event of type h at t. A model for nested case-control sampling is therefore given, via (8.2) and (8.4), by specifying, for each h, i and each t with $Y_{hi}(t) = 1$, the sampling distributions $\pi_t(\cdot | h, i)$ over sets $\mathbf{r}$ in $\mathcal{P}_i$.

Typically, if h indicates population strata, sampling will be done within these strata, i.e., with $\pi_t(\mathbf{r} | h, i)$ a distribution over sets of subjects in population stratum h at time t. But the sampling may be across strata with $\pi_t(\mathbf{r} | h, i) = \pi_t(\mathbf{r} | i)$ for all h. Such a strategy may be necessary if information on population strata is collected only on the sample or when the single stratum model is of primary interest and the multiple population strata are used for model checking. As examples one may adopt simple or stratified random sampling of controls within population strata or for all population strata combined.

To derive a partial likelihood, we introduce the counting processes

$$N_{(h,\mathbf{r})}(t) = \sum_{i \in \mathbf{r}} N_{(h,i,\mathbf{r})}(t),$$

counting the number of times an event of type h occurs in $[0, t]$ together with a sampled risk set equal to $\mathbf{r}$, and their intensity processes

$$\lambda_{(h,\mathbf{r})}(t) = \sum_{i \in \mathbf{r}} Y_{hi}(t)\alpha_{h0}(t)\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{hi}(t))\pi_t(\mathbf{r} | h, i; \boldsymbol{\beta}).$$

The intensity processes $\lambda_{(h,i,\mathbf{r})}$ may be factorized as

$$\lambda_{(h,i,\mathbf{r})}(t) = \lambda_{(h,\mathbf{r})}(t)\pi_t(i | h, \mathbf{r}; \boldsymbol{\beta}_0),$$

with

$$\pi_t(i \mid h, \mathbf{r}; \boldsymbol{\beta}_0) = \frac{Y_{hi}(t)\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{hi}(t))\pi_t(\mathbf{r}\mid h,i)}{\sum_{j\in\mathbf{r}} Y_{hj}(t)\exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{hj}(t))\pi_t(\mathbf{r}\mid h,j)}$$

being the conditional probability that the ith individual experiences an event of type h at t given $\mathcal{F}_{t-}$ and that an event of type h occurs among individuals in the set $\mathbf{r}$ at t.

Multiplying together these conditional probabilities over all observed events, we arrive at the partial likelihood

$$\mathcal{L}_\tau^{\mathbf{r}}(\boldsymbol{\beta}) = \prod_{u\in[0,\tau]}\prod_{h=1}^{k}\prod_{\mathbf{r}\in\mathcal{P}}\prod_{i\in\mathbf{r}}\left\{\frac{Y_{hi}(u)\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{hi}(u))\pi_u(\mathbf{r}\mid h,i)}{\sum_{j\in\mathbf{r}} Y_{hj}(u)\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{hj}(u))\pi_u(\mathbf{r}\mid h,j)}\right\}^{\Delta N_{(h,i,\mathbf{r})}(u)}. \quad (8.5)$$

As before the estimator $\widehat{\boldsymbol{\beta}}$, obtained by maximizing this partial likelihood, will be asymptotically multivariate normally distributed around the true value $\boldsymbol{\beta}_0$ with a covariance matrix that may be estimated by the inverse of the observed information matrix (evaluated at $\widehat{\boldsymbol{\beta}}$). Formal proofs may be written out along the lines of Section 6 provided that the regularity conditions stated there hold for each type.

Note that if h represent population strata and the sampling is done across these strata, if the ith subject fails, then $Y_{hj}(t_i) = 0$ for subjects not in the same population stratum as i and such subjects will not contribute to the partial likelihood. Thus, as one might expect, there may be a severe efficiency penalty for not sampling within population stratum.

To be able to estimate the type specific baseline intensities

$$A_{h0}(t) = \int_0^t \alpha_{h0}(u)\mathrm{d}u$$

we must assume that the sampling distributions $\pi_t(\cdot \mid h, i)$ are of the form (4.1) for each type h. Thus, assume that for each h and t, at which at least one individual is at risk for an event of type h, there exists, conditional on $\mathcal{F}_{t-}$, a sampling distribution $\pi_t(\cdot \mid h)$ over sets $\mathbf{r}$ in $\mathcal{P}$ such that

$$\pi_t(\mathbf{r}\mid h,i) = \frac{\pi_t(\mathbf{r}\mid i,h)}{p_t(i\mid h)} \quad (8.6)$$

for $i \in \mathbf{r}$, where

$$p_t(i\mid h) = \sum_{\mathbf{r}\in\mathcal{P}_i} \pi_t(\mathbf{r}\mid h). \quad (8.7)$$

is assumed to be positive for each i with $Y_{hi}(t) = 1$.

Then by the same argument as in Section 4, we obtain the estimators

$$\widehat{A}_{h0}(t,\widehat{\boldsymbol{\beta}}) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \frac{\mathrm{d}N_{(h,\mathbf{r})}(u)}{\sum_{j\in\mathbf{r}} Y_{hj}(u)\exp(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}_{hj}(u))/p_u(j\mid h)} \quad (8.8)$$

for the integrated baseline intensities. For fixed $\mathbf{Z}_{h0}$ one may estimate $A_h(t,\mathbf{Z}_{h0}) = \exp(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{h0})A_{h0}(t)$ by

$$\widehat{A}_h(t;\mathbf{Z}_{h0}) = \exp(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}_{h0})\widehat{A}_{h0}(t,\widehat{\boldsymbol{\beta}}),$$

and the asymptotic properties may be derived as in Section 6.

The integrated intensity estimators may form the basis for model checking procedures as for the full cohort model, see e.g. the review by Andersen *et al.* (1992, Section VII.3). Furthermore, for Markov process models, they may product-integrated to give estimators for transition probabilities as described Andersen *et al.* (1991) and Andersen *et al.* (1992, Section VII.2.3).

9 Aalens linear regression model

We have in this paper considered the proportional hazards regression model (2.2). An alternative to this model is Aalen's linear regression model, see the review in Andersen *et al.* (1992, Section VII.4.1) and the references therein. For this model it is assumed that the intensity process for the N_i are given by

$$\lambda_i(t) = Y_i(t)\{\beta_0(t) + \beta_1(t)Z_{i1}(t) + \cdots + \beta_p(t)Z_{ip}(t)\}, \tag{9.1}$$

where the β_j are arbitrary regression functions only restricted by the requirement that the λ_i should be non-negative.

In this section we will indicate how estimation of the integrated regression functions

$$B_j(t) = \int_0^t \beta_j(u)\mathrm{d}u; \qquad j = 1, \ldots p;$$

may be performed based on sampled cohort data. To this end we assume that the sampling distribution satisfies (4.1), and note that by (2.6) the $N_{(i,\mathbf{r})}$ then have intensity processes

$$\lambda_{(i,\mathbf{r})}(t) = \pi_t(\mathbf{r})\mathbf{Y}_i(t)\boldsymbol{\beta}(t) \tag{9.2}$$

with

$$\mathbf{Y}_i(t) = (1, Z_{i1}(t), \ldots, Z_{ip}(t)) \cdot Y_i(t)/p_t(i) \tag{9.3}$$

and

$$\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^{\mathsf{T}}.$$

We furthermore introduce the $|\mathbf{r}|$-dimensional column vector $\mathbf{N}_\mathbf{r}(t)$ with elements $N_{(l,\mathbf{r})}(t)$, $l \in \mathbf{r}$, and define $\boldsymbol{\lambda}_\mathbf{r}(t)$ and $\mathbf{M}_\mathbf{r}(t)$ similarly. We also introduce the $|\mathbf{r}| \times (p+1)$ dimensional matrix $\mathbf{Y}_\mathbf{r}(t)$ with rows $\mathbf{Y}_l(t)$, $l \in \mathbf{r}$, cf. (9.3).

As an estimator for $\mathbf{B}(t) = (B_0(t), B_1(t), \ldots, B_p(t))^{\mathsf{T}}$ we then propose

$$\hat{\mathbf{B}}(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J_\mathbf{r}(u)\mathbf{Y}_\mathbf{r}^-(u)\mathrm{d}\mathbf{N}_\mathbf{r}(u). \tag{9.4}$$

Here $\mathbf{Y}_\mathbf{r}^-(t)$ is a generalized inverse of $\mathbf{Y}_\mathbf{r}(t)$, i.e. a $(p+1) \times |\mathbf{r}|$ matrix satisfying

$$\mathbf{Y}_\mathbf{r}^-(t)\mathbf{Y}_\mathbf{r}(t) = \mathbf{I},$$

the $(p+1) \times (p+1)$ identity matrix, and

$$J_\mathbf{r}(t) = I(\mathrm{rank}\,\mathbf{Y}_\mathbf{r}(t) = p+1)$$

is the predictable indicator of $\mathbf{Y}_\mathbf{r}(t)$ having full rank. Note that this implies that each sampled risk set must include at least $p+1$ subjects.

The motivation for the estimator (9.4) is the following. By (9.2) we may write (c.f. (4.5))

$$\widehat{\mathbf{B}}(t) = \mathbf{B}^*(t) + \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J_\mathbf{r}(u)\mathbf{Y}_\mathbf{r}^-(u)\mathrm{d}\mathbf{M}_\mathbf{r}(u). \tag{9.5}$$

with

$$\mathbf{B}^*(t) = \int_0^t \left\{\sum_{\mathbf{r}\in\mathcal{P}} J_\mathbf{r}(u)\pi_u(\mathbf{r})\right\}\beta(u)\mathrm{d}u.$$

If $\mathbf{Y}_\mathbf{r}(t)$ has full rank with high probability then $\mathbf{B}^*(t)$ is almost the same as $\mathbf{B}(t)$ and (9.5) then gives that the estimator (9.4) is almost unbiased.

The variance of $\widehat{\mathbf{B}}(t)$ may be estimated by the optional variation process of $\widehat{\mathbf{B}} - \mathbf{B}^*$, i.e. by

$$\widehat{\mathbf{\Sigma}}(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} J_\mathbf{r}(u)\mathbf{Y}_\mathbf{r}^-(u)\mathrm{diag}\left(\mathrm{d}\mathbf{N}_\mathbf{r}(u)\right)\mathbf{Y}_\mathbf{r}(u)^\mathsf{T}, \tag{9.6}$$

where diag(a) for a vector a is the diagonal matrix with the elements of a in the diagonal.

To actually calculate $\widehat{\mathbf{B}}(t)$ and $\widehat{\mathbf{\Sigma}}(t)$ a choice of generalized inverse must be made. A simple possibility is

$$\mathbf{Y}_\mathbf{r}^-(t) = \left(\mathbf{Y}_\mathbf{r}(t)^\mathsf{T}\mathbf{Y}_\mathbf{r}(t)\right)^{-1}\mathbf{Y}_\mathbf{r}(t)^\mathsf{T}$$

corresponding to an unweighted least squares principle. A weighted least squares approach suggests the use of

$$\mathbf{Y}_\mathbf{r}^-(t) = \left(\mathbf{Y}_\mathbf{r}(t)^\mathsf{T}\widehat{\mathbf{W}}_\mathbf{r}(t)\mathbf{Y}_\mathbf{r}(t)\right)^{-1}\mathbf{Y}_\mathbf{r}(t)^\mathsf{T}\widehat{\mathbf{W}}_\mathbf{r}(t)$$

for a suitably chosen $|\mathbf{r}| \times |\mathbf{r}|$ diagonal weight matrix $\widehat{\mathbf{W}}_\mathbf{r}(t)$.

It should be possible to derive large sample properties of the estimator $\widehat{\mathbf{B}}(t)$ using the martingale representation (9.5) along the lines of the proofs for the full cohort; c.f. McKeague (1988), Huffer and McKeague (1991) as well as the summary in Andersen *et al.* (1992; Section VII.4.2). We will not go into this here, however. The practical applicability of the estimator (9.4) also needs to be investigated, and one may expect that $\widehat{\mathbf{B}}(t)$ based on the nonparametric model (9.1) will need much larger sampled risk sets to behave reasonably than what is the case for the estimators for the semiparametric model (2.2).

10 Discussion

The general framework we have presented makes it possible to analyze a large class of sampling designs. The three completely novel (classes of) designs given in Examples 5.3 - 5.5 illustrate the potential usefulness of the methods. Many techniques available for the analysis of full cohort data are accommodated with little change for sampled data. In this paper alone, we have given estimation methods for relative risks (using the partial

likelihood), baseline hazards, survival probabilities, and extensions to multistrata and multistate problems, and to the Aalen linear regression model. Estimation of relative mortality is developed in Borgan and Langholz (1993). Further, the marked point process framework can be generalized to accommodate other design problems. For instance, in Langholz and Borgan (1992), a simple generalization of the mark space described in Section 2 is used to derive a partial likelihood when failures are also to be sampled from the cohort.

In Section 5, we were careful to point out that the new designs and associated partial likelihood might, but not necessarily would, be more efficient than "standard" methods (by which we were thinking of simple nested case-control sampling). There are actually two aspects to the issue of efficiency. The first is the efficiency of the design (compared to the full cohort) when the most efficient analytic method is used. The second is the efficiency of the partial likelihood for the given design compared to the optimal method (for this design). It is important stress that the partial likelihood approach presented here provides *a* method for estimation of the regression parameters for a large class of sampling designs but, since optimal analytic methods for a given sampling design have not been characterized, it is an open question as to when the partial likelihood for a given design is efficient relative to the optimal. Greenwood and Wefelmeyer (1990) show the asymptotic efficiency of the partial likelihood for estimation in the full cohort case. Their approach should be readily adaptable to the sampled cohort situation and, in fact, has been used to show the efficiency of the partial likelihood for simple nested case-control sampling (Scheepker, 1992). In Section 6, we provide the basis for (large sample) comparisons of partial likelihood analyses of designs within this class. In particular, efficiency relative to the full cohort or to the simple nested case-control sampling is possible. For instance, using asymptotic variance formulas (7.5) and (7.7), the stratified sampling method of Section 5.3 was found to have much smaller asymptotic variance than simple nested case-control sampling in situations of practical importance (Langholz and Borgan, 1992).

Interestingly, up until now, simple nested case-control sampling has been the only sampling design for which analysis is based on a partial likelihood. Case-cohort sampling, perhaps the only other cohort sampling method which has been used for actual epidemiologic studies, relies on a "pseudo-likelihood" approach because, as shown in Example 5.6, the partial likelihood does not make use of non-subcohort failures. A similar phenomenon occurs with variants of nested case-control sampling in which subjects may serve as a control only once (Prentice, 1986*b*, Robins *et al.*, 1989). Designs II and III of Langholz and Thomas (1991) are examples of sampling designs which do not belong to the class we have considered. The path sets, used in the stratified sampling by Langholz and Thomas, are based on all failure times which occur over the study period and, thus, the sampling distributions are not predictable.

In earlier, work Goldstein and Langholz (1992) developed the asymptotic theory for (simple) nested case-control sampling based on a different model from that given here. In their model, just after a change in Y_i or N_i for some subject i in the cohort, a set of controls is randomly (and independently) sampled for each at risk subject. Then, when a failure occur, the sampled risk set would be already established. The counting processes then just count failure occurrences, as in the full cohort framework of Andersen and Gill (1982), and the fictitious sampling is predictable under an obvious enlarged filtration. In the marked point process approach of the present paper, the counting processes count joint failure *and* sampled risk set occurrences. The probability laws for the sampling are predictable but the sampling itself is adapted (but not predictable) with respect to the filtration $(\mathcal{F}_t)$. The observed scores from both models are identical but the score *process* (3.8) is exactly a

martingale while Goldstein and Langholz's is a martingale plus an additional term. This second term is due to the additional variation generated by the multiplicity of (fictitiously) sampled risk sets and is asymptotically negligible. The approach given here not only vastly simplifies proofs, allows for a partial likelihood interpretation, and leads, quite naturally, to the estimator of the cumulative baseline hazard but also reflects how nested case-control sampling is actually done. We note, for completeness, that the earlier approach of Goldstein and Langholz does generalize to accommodate some of the sampling schemes in the class considered here; Langholz and Goldstein (1992) developed the stratified sampling method in Example 5.3 in this way.

Acknowledgements

This research was initiated when Ørnulf Borgan and Bryan Langholz were on sabbatical leave at the MRC Biostatistics Unit, Cambridge, England, during the academic year 1991/92. The MRC Biostatistics Unit is acknowledged for its hospitality and for providing us with the best working facilities during this year. We acknowledge the support of Norwegian Research Council for Science and the Humanities (Ørnulf Borgan), United States National Science Foundation grant DMS 90-05833 (Larry Goldstein) , and United States National Cancer Institute grant CA14089 (Bryan Langholz).

References

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N., and Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics* **41**, 921–932.

Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12**, 97–158.

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

Andersen, P. K., Christensen, E., Fauerholdt, L., and Schlichting, P. (1983). Measuring prognosis using the proportional hazards model. *Scand. J. Statist.* **10**, 49–52.

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

Andersen, P. K., Hansen, L. S., and Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process. *Scand. J. Statist.* **18**, 153–167.

Arjas, E. (1989). Survival models and martingale dynamics (with discussion). *Scand. J. Statist.* **16**, 177–225.

Billingsley, P. (1961). *Statistical Inference for Markov Processes.* Univ. of Chicago Press, Chicago.

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

Brémaud, P. (1981). *Point Processes and Queues: Martingale Dynamics.* Springer-Verlag, New York.

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

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research.* Volume 2 – *The Design and Analysis of Cohort Studies, IARC Scientific Publications*, Vol. 82. International Agency for Research on Cancer, Lyon.

Breslow, N. E., Lubin, J. H., Marek, P., and Langholz, B. (1983). Multiplicative models and cohort analysis. *J. Amer. Statist. Assoc.* **78**, 1–12.

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

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

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

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* (to appear).

Greenwood, P. E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stoch. Proc. Appl.* **36**, 353–370.

Hjort, N. L. and Pollard, D. (1993). Simpler proofs and sharper results for asymptotic properties of the cox regression estimator and other minimisers of convex criteria. Statistical research report, Institute of Mathematics, University of Oslo. (to appear).

Huffer, F. W. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.* **86**, 114–129.

Karr, A. F. (1986). *Point Processes and Their Statistical Inference.* Marcel Dekker, New York.

Langholz, B. and Borgan, Ø. (1992). Stratified nested case-control sampling in the Cox regression model. Statistical Research Report 6/92, Institute of Mathematics, University of Oslo. (to appear in *Biometrika*).

Langholz, B. and Goldstein, L. (1992). Stratified nested case-control sampling in the cox regression model. Technical Report 34/92, Department of Preventive Medicine, Biostatistics Division, University of Southern California.

Langholz, B. and Thomas, D. C. (1991). Efficiency of cohort sampling designs: Some surprising results. *Biometrics* **47**, 1563–1571.

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

McKeague, I. W. (1988). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model. *Contemp. Math.* **80**, 139–152.

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

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

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

Prentice, R. L. and Self, S. G. (1983). Asymptotic distribution theory for Cox-type regressions models with general relative risk form. *Ann. Statist.* **11**, 804–813.

Robins, J. M., Gail, M. H., and Lubin, J. H. (1986). More on "Biased selection of controls for case-control analysis of cohort studies". *Biometrics* **42**, 293–299.

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

Scheepker, D. (1992). Asymptotic efficiency for nested case-control sampling. Manuscript, University of Southern California.

Schuster, E. F. and Sype, W. R. (1987). On the negative hypergeometric distribution. *International Journal of Mathematical Education in Science and Technology* **18**, 453–459.

Scott, A. and Wild, C. (1991). Fitting logistic models in stratified case-control studies. *Biometrics* **47**, 497–510.

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

Thomas, D. 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. *J. Roy. Statist. Soc. A* **140**, 469–491.

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