stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 89/82 OKTOBER
(DEPARTMENT OF MATHEMATICAL STATISTICS)

Ø. BORGAN & R.D. GILL

CASE-CONTROL STUDIES IN A MARKOV CHAIN SETTING

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: Primary: 62M02, 62M05
Secondary: 62P10

Case-control studies in a Markov chain setting *)

by

Ø. Borgan **) & R.D. Gill

ABSTRACT

For some types of retrospective case-control studies we show how non-parametric estimators of the forces of morbidity for chronic diseases may be constructed. For other types of study, estimators of closely related quantities may be also derived. In each case non-parametric testing can be carried out to investigate whether a certain factor or condition is a "risk-factor" for the disease. An application is given to the example discussed by AALEN et al. (1980).

KEY WORDS & PHRASES: *Case-control studies, retrospective studies, Markov process models, counting processes, Aalen-Nelson estimator, linear rank tests for censored data*

---

*) This report will be submitted for publication elsewhere.

**) Department of Statistics, Math. Inst., University of Oslo

1. INTRODUCTION

The purpose of many investigations in epidemiology and medicine is to determine whether a certain factor or condition is a "risk factor" for a specific disease. If the disease is a rare one, it is inconvenient to perform a prospective study, or a follow-up study, and one usually has to resort to a retrospective study, most often a case-control study. In such a study samples of individuals with and without the disease are selected, and one determines whether the individuals are exposed to the risk factor in question or not.

Historically case-control studies have roots back in the 19th century (see LILIENFELD & LILIENFELD, 1979), but a thorough investigation of the statistical theory for such studies started only 30 years ago with the important paper by CORNFIELD (1951). This paper, and later works of CORNFIELD (1956), MANTEL & HAENZEL (1959), and CORNFIELD & HAENZEL (1960), still constitute the philosophical basis for most works on case-control studies. The key idea in this theory is the following: it is an algebraic fact that the odds ratio of the risk factor among the diseased and non-diseased equals the odds ratio of the disease among those with and without the factor. Since the latter is close to the relative risk of the disease for rare diseases, while the former may be estimated from a case-control study, case-control studies may be used to investigate the importance of a possible risk factor.

Even if the basic idea is quite simple, a number of problems is of great importance in practical investigations. Important problems are how to select the cases and controls to avoid biased samples, and how to stratify or to "match" cases and controls to take possible heterogeneity in the population into account. These and other problems are much discussed in epidemiological circles, as can be seen from the recent issue on case-control studies of *The Journal of Chronic Diseases* (1979, vol. 32, no. 1/2).

During the last decade the classical methods in case-control studies have been further developed to allow the odds ratio to depend on concomitant information ("covariates"), giving regression type models. This new theory, as well as the classical one, is reviewed in the recent book on case-control studies by BRESLOW & DAY (1980), which contains a comprehensive

set of references.

In their famous paper of 1959 Mantel and Haenzel stated that "a primary goal is to reach the same conclusions in a retrospective (case-control) study as would have been obtained from a forward (follow-up) study, if one had been done". In our opinion a lot of work still has to be done before the goal of Mantel and Haenzel is achieved, our main reason being that the usual methods for the analysis of case-control studies only use *counts* of the various events and do not take the *timing* of the events into account as is usually done for prospective studies. Thus, the theory of case-control studies applies results from the theory of contingency tables and not from the theory of stochastic processes, as we feel should be the case.

It is the purpose of the present paper to make a first attempt at developing a theory for case-control studies for a homogeneous population in a Markov chain setting. The paper has been inspired by the recently developed non-parametric methods of AALEN (1978) and ANDERSEN et al. (1982), and takes the discussion of retrospective observational plans in HOEM (1969), AALEN et al. (1980), and BORGAN (1980) as its starting point. We show how one may derive non-parametric estimators for (quantities closely related to) the forces of morbidity for chronic diseases and how non-parametric testing may be carried out for some types of case-control studies. An application related to the example discussed by AALEN et al. (1980) is also given.

It turns out that the estimators and test statistics appropriate to many types of study all have the same structure. Thus we only give a detailed derivation of large sample properties for the first type considered; for the other types one can immediately write down the analogous properties. Moreover, in the body of the paper we only give a heuristic derivation of these properties. The technical problems of a rigorous derivation, which are surprisingly heavy ones, are deferred to Appendix 1.

It is easy to point out a number of shortcomings in the methods given in this paper. Since we are only considering Markov chain models with a finite state space, we have to assume that the risk factors only have a finite number of levels.

A more serious limitation in many applications (e.g. cancer research) is that Markov chain models do not take the duration of exposure to the

factor into account. Finally, we only consider methods applicable to a homogeneous population. However, these shortcomings are not specific to our methods. Moreover, even for situations where the duration of the exposure to the risk factor is considered important, one usually has a Markov chain model under the hypothesis that the factor has no effect on the development of the disease. Therefore, we believe that the tests suggested in this paper will catch most interesting deviations from this hypothesis in such situations too. Nevertheless, more work needs to be done before a satisfactory theory for case-control studies in a stochastic process setting is established. One should for instance try to develop a theory for semi-Markov models (here the paper of HOEM (1972) may be a starting point), and to discuss Markov chain models where the intensities may depend on certain covariates so as to model population heterogeneity.

2. THE MARKOV CHAIN MODELS

The Markov chain models to be considered in this paper may be exemplified by the simple model given in Fig. 1.

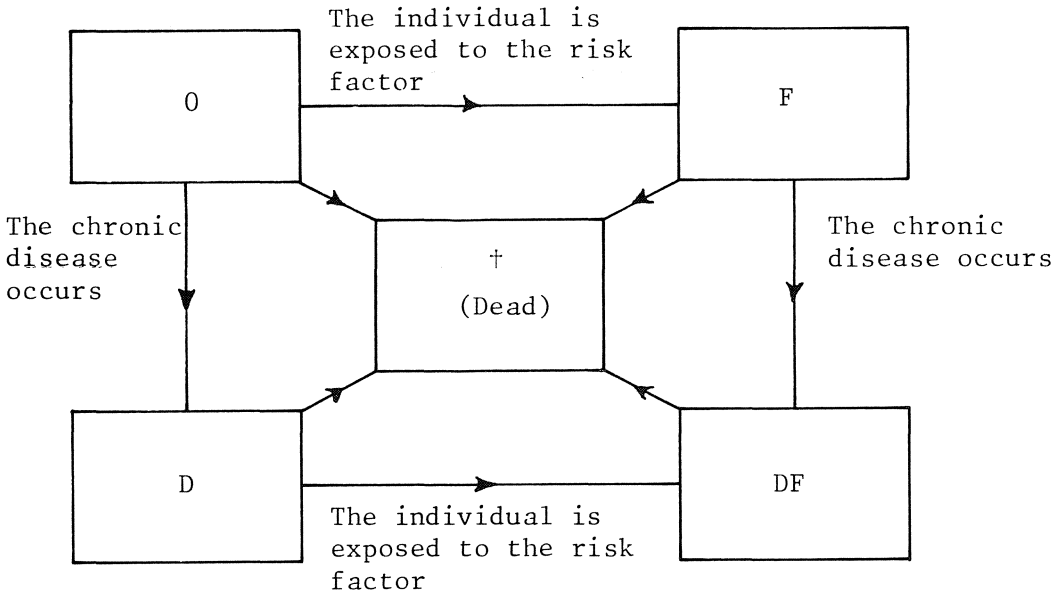

Fig. 1. A simple Markov chain model for the evaluation of a certain risk factor

In this model an individual may be healthy and not exposed to a certain risk factor F. Such an individual is in state 0. Then the individual may become exposed to the factor (the model in Fig. 1 assumes that this is an irreversible event) and move to state F. Thereafter the individual may get a certain chronic disease and go on to state DF. If the chronic disease occurs before the person is exposed to the risk factor, the individual takes the route $0 \to D \to DF$. A person may also die. This corresponds to a move to the state †.

More generally, in this paper we will consider a non-homogeneous, time continuous Markov chain model $\{S(x):x\geq 0\}$ with *finite* state space J, which may be written as a disjoint union $J = H \cup I \cup \mathcal{D}$. The subset H of the state space corresponds to various exposure statuses for healthy individuals, while I corresponds to the same statuses for diseased (ill) individuals. The subset $\mathcal{D}$ consists of "dead" states. Thus, in Fig. 1 we have $H = \{0,F\}$, $I = \{D,DF\}$, and $\mathcal{D} = \{\dagger\}$. It should be noted that the theory presented in this paper only assumes that $\mathcal{D}$ is an *absorbing* subset of states. Thus, the states in $\mathcal{D}$ may also represent some kind of censoring such as emigration. However, we will throughout the paper denote a transfer to a state in $\mathcal{D}$ as a death.

We will assume in most of the paper that the transition probabilities $P_{ij}(x,y) = P\{S(y) = j \mid S(x) = i\}$, $i,j \in J$, are absolutely continuous in (x,y), and that the intensities or forces of transition, defined as

$$(2.1) \qquad \alpha_{ij}(x) = \lim_{y \downarrow x} P_{ij}(x,y)/(y-x)$$

for $i,j \in J$, $i \neq j$, exist and are integrable. So *cumulative* (or *integrated*) intensities defined as

$$(2.2) \qquad A_{ij}(x) = \int_0^x \alpha_{ij}(y)dy \qquad i \neq j$$

also exist.

In fact *all* the results we give hold under the simple assumption that cumulative intensities exist; they need not be continuous, let alone absolutely continuous. For ease of exposition we have chosen to work in the special case of absolutely continuous cumulative intensities. In Appendix 2 we explain the more general model and point out the significance of this

generalization.

We introduce the notation $P_{i\mathcal{B}}(x,y) = \sum_{j\in\mathcal{B}} P_{ij}(x,y)$ for any subset $\mathcal{B}$ of J and $\alpha_{i\mathcal{B}}(x) = \sum_{j\in\mathcal{B}} \alpha_{ij}(x)$ for any subset $\mathcal{B}$ of $J\backslash\{i\}$. Since we only consider chronic diseases in this paper (or only the *first* occurrence of a disease), we assume throughout that $\alpha_{iH} \equiv 0$ for all $i \in I$. Of couse $\alpha_{ij} \equiv 0$ for all $i \in \mathcal{D}$, $j \in L = H \cup I$. Moreover, we will assume that no individual is diseased at birth and that the initial probability distribution at age zero is $\{\pi_h: h \in H\}$, i.e. $P\{S(0) = h\} = \pi_h$; $h \in H$; $\sum_{h\in H} \pi_h = 1$.

Now, in connection with the example of Fig. 1, the epidemiological problem is the following: is the factor F in fact a risk factor? In the present set-up we will say that this is the case if $\alpha_{F,DF}(x) \geq \alpha_{0D}(x)$ for all $x > 0$, with strict inequality for at least some x. Thus, our statistical problems are in general to estimate quantities related to the α_{hi}'s for $h \in H$, $i \in I$ and test hypotheses concerning these functions. These statistical problems can be equivalently formulated in terms of the A_{hi}'s. These problems will be considered in the succeeding paragraphs for various sampling frames for the cases and controls. Before we turn to that, we close this section with some results and notation which will be useful in the sequel.

The notation $\bar{P}_{ij}(x,y)$, $i,j \in L = H \cup I$, will be used for the transition probabilities of the partial Markov chain with state space L obtained by substituting zero for α_{ij} for all (i,j) with $j \in \mathcal{D}$ (HOEM, 1969). Moreover, we let $P_i(x) = P\{S(x) = i\} = \sum_{h\in H} \pi_h P_{hi}(0,x)$ for $i \in J$, $P_{\mathcal{B}}(x) = P\{S(x)\in\mathcal{B}\} = \sum_{i\in\mathcal{B}} P_i(x)$ for any subset $\mathcal{B}$ of J, and write $\bar{P}_i(x)$ and $\bar{P}_{\mathcal{B}}(x)$ for the similar partial quantities. If there is *non-differential mortality* for live individuals, i.e. $\alpha_{j\mathcal{D}} \equiv \mu$ for all $j \in L$, then

$$(2.3) \qquad P_{jL}(x,y) = \exp\left(-\int_x^y \mu(u)du\right),$$

and

$$(2.4) \qquad P_{ij}(x,y) = \bar{P}_{ij}(x,y)\exp\left(-\int_x^y \mu(u)du\right)$$

for all $i,j \in L$ (HOEM, 1969). It is seen that $\bar{P}_{ij}(x,y)$ in this situation (and only in this one, COHEN, 1972) is the conditional probability that an individual in state i at age x will be in state j at age $y \geq x$, given that

the individual is still alive at the latter age. If there is *non-differential mortality* for *healthy* individuals only, i.e. $\alpha_{h\mathcal{D}} \equiv \mu$ for all $h \in H$, then (2.4) is still valid for $i,j \in H$ (BORGAN, 1980).

3. PREVALENT CASES OF A GIVEN AGE

3A. Introduction

Let us assume in this section that the cases consist of a sample of n individuals in some homogeneous population suffering from the chronic disease in question at a given age ζ. One collects a retrospective account of each individual's exposure and disease history. (In this paper, we choose to disregard all problems concerning the reliability of the information collected in retrospective studies.) We assume that the sample of cases is "representative" in the sense that the observations (from the cases) may be considered as independent, identically distributed realizations of the Markov chain with state space $L = H \cup I$ obtained from the one in Section 2 by conditioning on being in I at age ζ (see HOEM, 1969, for details). This Markov chain has transition probabilities

$$
\begin{aligned}
Q_{ij}(x,y) &= P\{S(y) = j \mid S(x) = i,\ S(\zeta) \in I\} \\
&= P_{ij}(x,y)P_{jI}(y,\zeta)/P_{iI}(x,\zeta)
\end{aligned}
\tag{3.1}
$$

for $i,j \in L$, $x \le y \le \zeta$. The transition intensities of the chain are

$$
\nu_{ij}(x) = \alpha_{ij}(x)P_{jI}(x,\zeta)/P_{iI}(x,\zeta) \tag{3.2}
$$

for $i,j \in L$, $i \neq j$, and $x < \zeta$.

Let $N_{hi}^{(n)}(x)$ be $\frac{1}{n}$ times the number of transitions directly from state $h \in H$ to state $i \in I$ reported by the n individuals in the age interval $[0,x]$, with $x \le \zeta$. Equivalently, $N_{hi}^{(n)}(x)$ is the relative frequency in the sample of the event "a direct transition from h to i at or before age x". Then $EN_{hi}^{(n)}(x) = \int_0^x \nu_{hi}(s)Q_h(s)ds$, where

$$Q_h(x) = P\{S(x) = h \mid S(\zeta) \in I\} = P_h(x)P_{hI}(x,\zeta)/P_I(\zeta).$$

By (3.2) this reduces to

$$(3.3) \qquad E\mathrm{N}_{hi}^{(n)}(x) = \int_0^x \alpha_{hi}(s)P_h(s)P_{iI}(s,\zeta)ds/P_I(\zeta).$$

Without any assumptions on the mortality in the various "live states" this expression is of little use. However, we get nice results when (i) there is *non-differential mortality* for *all live* individuals, i.e. $\alpha_{j\mathcal{D}} \equiv \mu$ for all $j \in L$, or (ii) when there is *non-differential mortality* for *healthy* individuals and *non-differential mortaility* for *diseased* individuals, i.e. $\alpha_{h\mathcal{D}} = \mu_0$ for $h \in H$ and $\alpha_{i\mathcal{D}} \equiv \mu_1$ for $i \in I$, and the disease is a rare one. We will in Subsections 3.B,C, and D give a careful treatment of (i) and return to (ii) in Subsection 3.E.

3.B. Non-differential mortality

When mortality is non-differential for all live individuals (2.3) and (3.3) give

$$(3.4) \qquad E\mathrm{N}_{hi}^{(n)}(x) = \int_0^x \alpha_{hi}(s)\bar{P}_h(s)ds/\bar{P}_I(\zeta),$$

so that

$$(3.5) \qquad A_{hi}(x) = \int_0^x \alpha_{hi}(s)ds = \bar{P}_I(\zeta)\int_0^x \frac{dE\mathrm{N}_{hi}^{(n)}(s)}{\bar{P}_h(s)}.$$

Since $\bar{P}_j(x)$ in the present context (cf. the end of Section 2) is the conditional probability of being in state j at age x for an individual who is still alive at age ζ, (3.5) suggests that one should draw the sample of m controls among *all* individuals alive at age ζ, not only among the non-diseased (as also pointed out by MIETTINEN, 1976). We will assume that the control individuals are sampled independently of the cases, and that they form a "representative" sample of the individuals alive in the population at age ζ, in the sense that the observations (from the controls) may be considered as independent, identically distributed realizations of the

Markov chain with state space L, initial distribution $\{\pi_h\}$, and transition intensities $\bar{P}_{ij}(x,y)$.

Let $Y_h^{(m)}(x)$ denote $\frac{1}{m}$ times the number of control individuals who report that they were in state h "just before" age x (i.e. $Y_h^{(m)}(\cdot)$ is left-continuous). Then $\bar{P}_h(x)$ may be estimated by $Y_h^{(m)}(x)$; the *prevalence* of the disease at age ζ, $\bar{P}_I(\zeta)$, may be estimated by $Y_I^{(m)}(\zeta) = \sum_{j\in I} Y_j^{(m)}(\zeta)$; and consequently an estimator for (3.5) is

$$\hat{A}_{hi}(x) = Y_I^{(m)}(\zeta)\int_0^x \frac{dN_{hi}^{(n)}(s)}{Y_h^{(m)}(s)} \quad .$$

The properties of this estimator may be derived using the theory which we present in the following pages. In most applications, however, the studied disease will be a rare one, which means that $P\{Y_I^{(m)}(\zeta) = 0\}$ will be close to unity. In such situations one will have to be content with the estimator

$$\hat{B}_{hi}(x) = \int_0^x \frac{dN_{hi}^{(n)}(s)}{Y_h^{(m)}(s)} \tag{3.6}$$

for $B_{hi}(x) = A_{hi}(x)/\bar{P}_I(\zeta)$. (We avoid symbols like $\hat{B}_{hi}^{(n,m)}(x)$ in order not to overburden the notation.)

3.C. Nonparametric estimation

We now give a heuristic derivation of the large sample properties of the estimators $\hat{B}_{hi}(x)$: $h \in H$, $i \in I$, $x \in [0,\zeta]$ given by (3.6). The main ingredients of a rigorous proof (using the techniques of BRESLOW & CROWLEY, 1974) are presented in Appendix 1.B. The heuristics are important because they lead to the correct expression for the asymptotic distribution of the (multivariate) stochastic process $\{\hat{B}_{hi}(\cdot): h \in H, i \in I\}$, which we denote simply by $\{\hat{B}_{hi}\}$, in the shortest possible way. Both the heuristics here and the formal proof in Appendix 1.B are prototypes for the large sample analysis of all the other statistical quantities we shall be considering.

We consider the process $\{\hat{B}_{hi}\}$ as a function of the processes $\{N_{hi}^{(n)}\}$ and $\{Y_h^{(m)}\}$. The latter are sample averages, and the function applied to their expected values $\{EN_{hi}^{(n)}\}$ and $\{EY_h^{(m)}\}$ yields the quantities being estimated, $\{B_{hi}\}$ (cf. below). We now imitate Rao's δ-method by approximating the

function with its first order Taylor expansion about these expected values. (The expansion is known in this context as the first order von Mises expansion, see SERFLING, 1980 Chap.6, or BOOS & SERFLING, 1980). This gives an approximate expression for $\{n^{\frac{1}{2}}(\hat{B}_{hi}-B_{hi})\}$ which is linear in $\{n^{\frac{1}{2}}(N_{hi}^{(n)}-EN_{hi}^{(n)})\}$ and $\{m^{\frac{1}{2}}(Y_h^{(m)}-EY_h^{(m)})\}$. Substituting the asymptotic Gaussian distributions of these processes should give the asymptotic Gaussian distribution of $\{n^{\frac{1}{2}}(\hat{B}_{hi}-B_{hi})\}$.

We will first have a brief look at $\{N_{hi}^{(n)}\}$ and $\{Y_h^{(m)}\}$. Let us write

$$(3.7)\qquad \begin{cases} EN_{hi}(t) = EN_{hi}^{(n)}(t) = \int_0^t \bar{P}_h(s)\,dB_{hi}(s) & \text{and} \\ EY_h(t) = EY_h^{(m)}(t) = \bar{P}_h(t). \end{cases}$$

Since there can be, for one individual, at most one transition directly from H to I, $\{N_{hi}^{(n)}\}$ is quite simply the empirical joint distribution function of the time of this transition and the pair of states $(h,i) \in H \times I$ involved. Thus $\{n^{\frac{1}{2}}(N_{hi}^{(n)}-EN_{hi})\}$ converges in distribution as $n \to \infty$ to a zero mean Brownian-bridge type multivariate Gaussian process $\{U_{hi}\}$ say, with covariance functions (equal to those of $\{N_{hi}^{(n)}\}$ with $n = 1$) given by

$$(3.8)\qquad \operatorname{cov}(U_{hi}(s),U_{k\ell}(t)) = \delta_{hi,k\ell}EN_{hi}(s\wedge t) - EN_{hi}(s)EN_{k\ell}(t),$$

where $\delta_{hi,k\ell}$ is a Kronecker delta and $\wedge$ denotes minimum. One control individual may enter and leave a state $h \in H$ any number of times. So weak convergence of $\{m^{\frac{1}{2}}(Y_h^{(m)}-EY_h)\}$ must be established directly. In Appendix 1.A we show that (precisely under the assumption that integrated intensities exist) this process converges in distribution as $m \to \infty$ to a multivariate Gaussian process $\{V_h\}$ with zero mean and with covariance functions (equal to those of $\{Y_h^{(m)}\}$ with $m = 1$) given by

$$\operatorname{cov}(V_h(s),V_k(t)) = \bar{P}_h(s)\bar{P}_{hk}(s,t) - \bar{P}_h(s)\bar{P}_k(t)$$

where we define $\bar{P}_{ij}(s,t)$ for $s > t$ as the "backwards" transition probability

of being in state j at time t given that the individual will be in state i at time s.

We are now ready to look at $\{\hat{B}_{hi}\}$. By (3.7) we have

$$B_{hi}(t) = \int_0^t (EY_h(s))^{-1} dEN_{hi}(s),$$

so that

$$(3.10)\qquad n^{\frac{1}{2}}(\hat{B}_{hi}(\cdot)-B_{hi}(\cdot)) = n^{\frac{1}{2}}\left(\int_0^{\cdot} \frac{dN_{hi}^{(n)}}{Y_h^{(m)}} - \int_0^{\cdot} \frac{dEN_{hi}}{EY_h}\right)$$

$$\approx n^{\frac{1}{2}}\left(\int_0^{\cdot} \frac{d(N_{hi}^{(n)}-EN_{hi})}{EY_h} - \int_0^{\cdot} \frac{(Y_h^{(m)}-EY_h)}{(EY_h)^2}\, dEN_{hi}\right)$$

$$= \int_0^{\cdot} \frac{d(n^{\frac{1}{2}}(N_{hi}^{(n)}-EN_{hi}))}{EY_h} - \left(\frac{n}{m}\right)^{\frac{1}{2}} \int_0^{\cdot} \frac{m^{\frac{1}{2}}(Y_h^{(m)}-EY_h)}{EY_h}\, dB_{hi}.$$

So we expect that, as $n,m \to \infty$ in such a way that $n/m \to \lambda \in [0,\infty)$, we shall have (jointly in $(h,i) \in H \times I$)

$$(3.11)\qquad n^{\frac{1}{2}}(\hat{B}_{hi}(\cdot)-B_{hi}(\cdot)) \xrightarrow{\mathcal{D}} \int_0^{\cdot} \frac{dU_{hi}}{EY_h} - \lambda^{\frac{1}{2}} \int_0^{\cdot} \frac{V_h}{EY_h}\, dB_{hi},$$

where $\{U_{hi}\}$ is independent of $\{V_h\}$ and each has the distribution derived above.

Two remarks are in order here: firstly, it turns out by the analysis in Appendix 1.B that we do have this convergence in distribution provided that $EY_h(x) \geq c > 0$ for all $x \in [0,\zeta]$ and all $h \in H$. (Clearly some assumption is needed to avoid difficulties with the integrands in the above expressions.) If the condition as it stands is not satisfied, we still have analogous results for a time interval $[x_1,x_2] \subset [0,\zeta]$ and a subset of states $h \in H$ for which the condition does hold. Secondly, the integral with respect to U_{hi} cannot be interpreted as an ordinary Lebesgue integral for each sample point. It must instead be interpreted (equivalently) either as a stochastic integral in the sense of MEYER (1976), or as an L_2 integral in the sense of DOOB (1953), or by formal integration by parts. However the

result is a Gaussian process whose covariance functions can be calculated in the natural way, as follows.

The asymptotic covariance functions of $\{n^{\frac{1}{2}}(\hat{B}_{hi}-B_{hi}\}$ are clearly the sum of components due to $\{U_{hi}\}$ and $\{V_h\}$. For the contribution due to $\{U_{hi}\}$, we use the relation

$$\text{(3.12)} \qquad \operatorname{cov}\left(\int_0^x \frac{dU_{hi}}{EY_h}, \int_0^y \frac{dU_{k\ell}}{EY_k}\right) = \int_{(u,v)\in[0,x]\times[0,y]} \frac{d(\operatorname{cov}(U_{hi}(u),U_{k\ell}(v))}{EY_h(u)EY_k(v)} .$$

(Integrating with respect to the signed measure generated by the bivariate "distribution function" $\operatorname{cov}(U_{hi}(\cdot),U_{k\ell}(\cdot))$ given in (3.9).) Now $EN_{hi}(x\wedge y)$ is the joint cumulative distribution function of the pair of random variables (X,Y), where $X \equiv Y$ has distribution function $EN_{hi}(x)$. Similarly $EN_{hi}(x)EN_{k\ell}(y)$ is the joint distribution function of the pair (X,Y), where X and Y are independent with distribution functions $EN_{hi}(x)$ and $EN_{k\ell}(y)$. Thus substituting (3.9) into (3.12), we find that the covariance in (3.12) is equal to

$$\delta_{hi,k\ell}\int_0^{x\wedge y} \frac{dEN_{hi}(u)}{(EY_h(u))^2} - \int_0^x \frac{dEN_{hi}(u)}{EY_h(u)} \int_0^y \frac{dEN_{k\ell}(v)}{EY_k(v)}$$

$$= \delta_{hi,k\ell}\int_0^{x\wedge y} \frac{dB_{hi}}{EY_h} - B_{hi}(x)B_{k\ell}(y).$$

For the contribution due to $\{V_h\}$, (3.7) and (3.9) give

$$\text{(3.14)} \qquad \operatorname{cov}\left(\lambda^{\frac{1}{2}}\int_0^x \frac{V_h}{EY_h}\, dB_{hi}, \lambda^{\frac{1}{2}}\int_0^y \frac{V_k}{EY_k}\, dB_{k\ell}\right) =$$

$$= \lambda \int_{u=0}^x \int_{v=0}^y \frac{\operatorname{cov}(V_h(u),V_k(v))}{EY_h(u)EY_k(v)}\, dB_{hi}(u)dB_{k\ell}(v)$$

$$= \lambda \int_{u=0}^x \int_{v=0}^y \frac{\bar{P}_h(u)\bar{P}_{hk}(u,v)}{\bar{P}_h(u)\bar{P}_k(v)}\, dB_{hi}(u)dB_{k\ell}(v) - \lambda B_{hi}(x)B_{k\ell}(y).$$

Adding these terms gives an expression for the limiting covariance functions of $\{n^{\frac{1}{2}}(\hat{B}_{hi}-B_{hi})\}$.

To be able to write down the natural estimators of these covariances we let $Y_{hk}^{(m)}(s,t)$ be 1/m times the number of control individuals who were in state h at time s and state k at time t, both for $s \le t$ and $s > t$, and write

$$EY_{hk}(s,t) = EY_{hk}^{(m)}(s,t) = \bar{P}_h(s)\bar{P}_{hk}(s,t).$$

By the results of Appendix 1 we can then state the following theorem.

THEOREM 3.1. *Suppose* $m,n \to \infty$ *in such a way that* $n/m \to \lambda \in [0,\infty)$ *and suppose that* $P_h(x) \ge c > 0$ *for all* $x \in [0,\tau]$ *and all* $h \in H$. *Then* $\{n^{\frac{1}{2}}(\hat{B}_{hi}-B_{hi}):(h,i) \in H \times I\}$ *given by* (3.6) *converges weakly in* $(D[0,\tau])^r$, *where* r *is the number of elements of* $H \times I$, *to a zero mean multivariate Gaussian process* $\{W_{hi}\}$, *with covariance structure given by*

$$\text{(3.15)}\qquad \operatorname{cov}(W_{hi}(x),W_{k\ell}(y)) = \delta_{hi,k\ell}\int_0^{x\wedge y}\frac{dB_{hi}(u)}{\bar{P}_h(u)} - (1+\lambda)B_{hi}(x)B_{k\ell}(y) + \lambda\int_{u=0}^{x}\int_{v=0}^{y}\frac{\bar{P}_h(u)\bar{P}_{hk}(u,v)}{\bar{P}_h(u)\bar{P}_k(v)}\,dB_{hi}(u)\,dB_{k\ell}(v).$$

This asymptotic covariance can be uniformly consistently estimated by substituting $\hat{B}_{hi}(u)$ *for* $B_{hi}(u)$, $Y_{hk}^{(m)}(u,v)$ *for* $\bar{P}_h(u)\bar{P}_{hk}(u,v)$, $Y_h^{(m)}(u)$ *for* $\bar{P}_h(u)$, *and* n/m *for* λ.

REMARK 3.1. Suppose the $\bar{P}_h$'s are known functions, so that one could consider estimating $B_{hi}(\cdot)$ by

$$\int_0^{\cdot}\frac{dN_{hi}^{(n)}}{\bar{P}_h}.$$

Then Theorem 3.1 gives the asymptotic behaviour of these estimators if we substitute $\lambda = 0$ (i.e. $m=\infty$) in the expression for the asymptotic covariance structure. Note that estimation of this structure is very much simpler in this case.

REMARK 3.2. Theorem 3.1 gives directly a weak uniform consistency result for $\hat{B}_{hi}$. Strong uniform consistency can also be proved using some of the techniques of Appendix 1.C.

3.D. Nonparametric testing

We now turn to the nonparametric testing problem. Suppose we want to test the hypothesis H_0 that for a given set $\mathcal{R} \subset \mathcal{H} \times \mathcal{I}$, $B_{hi} = B$ for all $(h,i) \in \mathcal{R}$ for some unknown integrated intensity B. Note that the hypothesis states equivalently that A_{hi} is the same for all $(h,i) \in \mathcal{R}$. Following the line of thought in ANDERSEN et al. (1982), we will base a test for this hypothesis on the vector of statistics $\hat{Z} = (Z_{hi}(\zeta), (h,i) \in \mathcal{R})$, given by

$$(3.16) \qquad Z_{hi}(x) = n^{\frac{1}{2}}\Big(\int_0^x L(Y_\bullet^{(m)}(u))\,dN_{hi}^{(n)}(u) - \int_0^x L(Y_\bullet^{(m)}(u))\,\frac{Y_h^{(m)}(u)}{Y_\bullet^{(m)}(u)}\,dN_\bullet^{(n)}(u)\Big) =$$

$$= n^{\frac{1}{2}} \sum_{(k,\ell)\in\mathcal{R}} \int_0^x L(Y_\bullet^{(m)}(u))\Big(\delta_{hi,k\ell} - \frac{Y_h^{(m)}(u)}{Y_\bullet^{(m)}(u)}\Big)dN_{k\ell}^{(n)}$$

where L is some fixed function, $N_\bullet^{(n)} = \sum_{(h,i)\in\mathcal{R}} N_{hi}^{(n)}$ and $Y_\bullet^{(m)} = \sum_{(h,i)\in\mathcal{R}} Y_h^{(m)}$. Note that since in general some of the summands in the expression for $Y_\bullet^{(m)}$ may be identical, we do not necessarily have $Y_\bullet^{(m)} \leq 1$.

The special choices $L(y) = y$ and $L(y) = 1$ give test statistics of the Gehan-Breslow type and of the log-rank type respectively (cf. ANDERSEN et al. 1982).

We shall have to assume in general that L has a continuous derivative L' in $(c_1-\varepsilon, c_2+\varepsilon)$ for some $\varepsilon > 0$ where c_1 and c_2 must satisfy the weak condition

$$0 < c_1 \leq EY_\bullet^{(m)}(x) \leq c_2$$

for all $x \in [0,\zeta]$. (Of course the differentiability condition is trivially satisfied for the L-functions mentioned above.) As in the estimation problem, even if this condition does not hold, analogous results can be given for an interval $[x_1,x_2] \subset [0,\zeta]$ on which it does hold. Also as before, we write $EY_\bullet$ for the function $EY_\bullet^{(m)}$, etc. When we replace the processes $Y_h^{(m)}$ and $N_{hi}^{(n)}$ in (3.16) with their expected values, we obtain by (3.7), under H_0,

the value zero. Thus under H_0 we have the following von Mises expansion:

$$Z_{hi}(\cdot) \approx \sum_{(k,\ell)\in R} \int_0^{\cdot} L(EY_\cdot)(\delta_{hi,k\ell} - \frac{EY_h}{EY_\cdot})d(n^{\frac{1}{2}}(N_{k\ell}^{(n)} - EN_{k\ell}))$$

$$+ (\frac{n}{m})^{\frac{1}{2}} \sum_{(k,\ell)\in R} \int_0^{\cdot} L(EY_\cdot)[-m^{\frac{1}{2}} \frac{(Y_h^{(m)} - EY_h)}{EY_\cdot} + \frac{EY_h}{(EY_\cdot)^2} m^{\frac{1}{2}}(Y_\cdot^{(m)} - EY_\cdot)]dEN_{k\ell}$$

$$+ (\frac{n}{m})^{\frac{1}{2}} \sum_{(k,\ell)\in R} \int_0^{\cdot} L'(EY_\cdot)m^{\frac{1}{2}}(Y_\cdot^{(m)} - EY_\cdot)\left(\delta_{hi,k\ell} - \frac{EY_h}{EY_\cdot}\right)dEN_{k\ell}$$

$$= \int_0^{\cdot} L(EY_\cdot)d(n^{\frac{1}{2}}(N_{hi}^{(n)} - EN_{hi})) - \int_0^{\cdot} L(EY_\cdot)\frac{EY_h}{EY_\cdot} d(n^{\frac{1}{2}}(N_\cdot^{(n)} - EN_\cdot))$$

$$+ (\frac{n}{m})^{\frac{1}{2}} \int_0^{\cdot} L(EY_\cdot)[-m^{\frac{1}{2}}(Y_h^{(m)} - EY_h) + \frac{EY_h}{EY_\cdot} m^{\frac{1}{2}}(Y_\cdot^{(m)} - EY_\cdot)]dB,$$

since the third term is identically zero, and since $dEN_\cdot = EY_\cdot dB$. We expect therefore that as $n,m \to \infty$ in such a way that $n/m \to \lambda \in [0,\infty)$, we shall have

$$\{Z_{hi}(\cdot)\} \xrightarrow{\mathcal{D}} \left\{ \int_0^{\cdot} L(EY_\cdot)dU_{hi} - \int_0^{\cdot} L(EY_\cdot)\frac{EY_h}{EY_\cdot} dU_\cdot \right.$$

$$\left. + \lambda^{\frac{1}{2}} \int_0^{\cdot} L(EY_\cdot)(-V_h + \frac{EY_h}{EY_\cdot} V_\cdot)dB \right\},$$

where $\{U_{hi}\}$ and $\{V_h\}$ are the Gaussian processes described in Section 3.B, $U_\cdot = \sum_{(h,i)\in R} U_{hi}$ and $V_\cdot = \sum_{(h,i)\in R} V_h$. Exactly as in that section we compute the two components of the asymptotic covariance structure of $\{Z_{hi}\}$ under H_0; the contribution to the asymptotic covariance of $Z_{hi}(x)$ and $Z_{k\ell}(y)$ due to $\{U_{hi}\}$ by (3.8) is

$$\delta_{hi,k\ell} \int_0^{x\wedge y} (L(EY_\cdot))^2 dEN_{hi} - \int_0^{x\wedge y} (L(EY_\cdot))^2 \frac{EY_h}{EY_\cdot} dEN_{k\ell}$$

$$- \int_0^{x\wedge y} (L(EY_\cdot))^2 \frac{EY_k}{EY_\cdot} dEN_{hi} + \int_0^{x\wedge y} (L(EY_\cdot))^2 \frac{EY_h}{EY_\cdot}\frac{EY_k}{EY_\cdot} dEN.$$

$$+\left(\int_0^x L(EY_\cdot)dEN_{hi} - \int_0^x L(EY_\cdot)\frac{EY_h}{EY_\cdot}dEN_\cdot\right)\cdot$$

$$\left(\int_0^y L(EY_\cdot)dEN_{k\ell} - \int_0^y L(EY_\cdot)\frac{EY_k}{EY_\cdot}dEN_\cdot\right)$$

$$= \int_0^{x\wedge y} (L(EY_\cdot))^2\left(\delta_{hi,h\ell} - \frac{EY_h}{EY_\cdot}\right)EY_k dB.$$

Similarly $\{V_h\}$'s contribution, by (3.9), is

$$\lambda \int_{u=0}^x \int_{v=0}^y L(EY_\cdot(u))L(EY_\cdot(v))[\bar{P}_h(u)\bar{P}_{hk}(u,v) - \frac{EY_h(u)}{EY_\cdot(u)}\bar{P}_\cdot(u)\bar{P}_{\cdot k}(u,v)$$

$$- \frac{EY_k(v)}{EY_\cdot(v)}\bar{P}_h(u)\bar{P}_{h\cdot}(u,v) + \frac{EY_h(u)}{EY_\cdot(u)}\frac{EY_k(v)}{EY_\cdot(v)}\bar{P}_\cdot(u)\bar{P}_{\cdot\cdot}(u,v)]dB(u)dB(v)$$

$$+ \lambda\left(\int_0^x L(EY_\cdot)(-\bar{P}_h + \frac{EY_h}{EY_\cdot}\bar{P}_\cdot)dB\right)\left(\int_0^y L(EY_\cdot)(-\bar{P}_k + \frac{EY_k}{EY_\cdot}\bar{P}_\cdot)dB\right)$$

$$= \lambda \int_{u=0}^x \int_{v=0}^y L(EY_\cdot(u))L(EY_\cdot(v))[EY_{hk}(u,v) - \frac{EY_h(u)}{EY_\cdot(u)}EY_{\cdot k}(u,v)$$

$$- \frac{EY_k(v)}{EY_\cdot(v)}EY_{\cdot h}(v,u) + \frac{EY_h(u)}{EY_\cdot(u)}\frac{EY_k(v)}{EY_\cdot(v)}EY_{\cdot\cdot}(u,v)]dB(u)dB(v).$$

Rather than write down the analogue to Theorem 3.1, we go one step further and formulate a theorem on the asymptotic distribution of a test statistic based on the row vector $\hat{Z} = (Z_{hi}(\zeta): (h,i) \in R)$. Let $\hat{C}$ be the natural estimator of the asymptotic null-hypothesis covariance matrix of $\hat{Z}$,

$$(3.17)\qquad \hat{C}_{hi,k\ell} = \int_{u=0}^\zeta (L(Y_\cdot^{(m)}(u)))^2\left(\delta_{hi,k\ell} - \frac{Y_h^{(m)}(u)}{Y_\cdot^{(m)}(u)}\right)Y_k^{(m)}(v)\frac{dN_\cdot^{(n)}(u)}{Y_\cdot^{(m)}(u)}$$

$$+ \frac{n}{m}\int_{u=0}^\zeta \int_{v=0}^\zeta L(Y_\cdot^{(m)}(u))L(Y_\cdot^{(m)}(v))[Y_{hk}^{(m)}(u,v) - \frac{Y_h^{(m)}(u)}{Y_\cdot^{(m)}(u)}Y_{\cdot k}^{(m)}(u,v)$$

$$- \frac{Y_k^{(m)}(v)}{Y_\cdot^{(m)}(v)} Y_{\cdot h}^{(m)}(v,u) + \frac{Y_h^{(m)}(u)Y_k^{(m)}(v)}{Y_\cdot^{(m)}(u)Y_\cdot^{(m)}(v)} Y_{\cdot\cdot}^{(m)}(u,v)\,] \frac{dN_\cdot^{(n)}(u)\,dN_\cdot^{(n)}(v)}{Y_\cdot^{(m)}(u)Y_\cdot^{(m)}(v)} .$$

This covariance matrix has rank at most equal to one less than the number of elements of R (the sum of all its elements is identically zero). In the following theorem, $\hat{C}^-$ denotes a generalized inverse of $\hat{C}$, and $\hat{Z}^T$ is the transpose of the vector $\hat{Z}$.

THEOREM 3.2. *Suppose that* $m,n \to \infty$ *in such a way that* $n/m \to \lambda \in [0,\infty)$. *Suppose that* L *is continuously differentiable on* $(c_1-\varepsilon, c_2+\varepsilon)$ *where* $\varepsilon > 0$ *and* $0 < c_1 \le EY_\cdot(x) \le c_2$ *for all* $x \in [0,\tau]$. *Suppose that* $EY_h(x) > 0$ *for all* $(h,i) \in R$ *and for all* x *in a subset of* $[0,\tau]$ *of positive* dB*-measure. Then under* H_0, *the test statistic* $\hat{Z}\hat{C}^-\hat{Z}^T$ *is asymptotically chi-squared distributed with number of degrees of freedom equal to one less than the number of elements of* R.

REMARK 3.3. A proof of Theorem 3.2 can be based on the results of Appendix 1. The condition on positivity of EY_h ensures that the asymptotic covariance matrix of $\hat{Z}$ has fullest possible rank. Intuitively speaking, the condition requires that there is some subinterval of $[0,\tau]$ during which all states h corresponding to $(h,i) \in R$ are occupied by some individuals in the control group, and during which some of the cases make a transition from h to i, $(h,i) \in R$.

REMARK 3.4. The estimator (3.17) will clearly be tedious to compute, since the second term is a double integral with a contribution at (u,v) whenever a transition in R occurs at both time u and time v. In some situations the test statistic simplifies considerably:

(i) In a number of practical applications each state $h \in H$ is represented once, and once only, together with a state $i \in I$ in R. Moreover, the disease is a rare one so that $\bar{P}_H(x)$ is close to unity for all $x \in [0,\tau]$, and hence $Y_\cdot^{(m)}(x)$ too. (If the control sample consists entirely of individuals in H at age τ we shall have $Y_\cdot^{(m)} \equiv 1$.) Then for *all* choices of L the statistics $\hat{Z}$ and $\hat{C}$ simplify to (setting $L(1) = 1$)

$$(3.18) \qquad Z_{hi}(\tau) = n^{\frac{1}{2}}\left(N_{hi}^{(n)}(\tau) - \int_0^\tau Y_h^{(m)}(u)\,dN_\cdot^{(n)}(u)\right)$$

and

$$(3.19)\qquad \hat{C}_{hi,k\ell} = \int_0^\zeta (\delta_{hi,k\ell} - Y_h^{(m)}(u)) Y_k^{(m)}(u)\, dN_\cdot^{(n)}(u)$$

$$+ \frac{n}{m} \int_0^\zeta \int_0^\zeta (Y_{hk}^{(m)}(u,v) - Y_h^{(m)}(u) Y_k^{(m)}(v))\, dN_\cdot^{(n)}(u)\, dN_\cdot^{(n)}(v).$$

(ii) If λ is small, i.e. the control sample is relatively large compared to the sample of cases, the second term of (3.17) is of less importance. The first term only consists of a single integral. Moreover, the first term has exactly the same form as in the situation described by ANDERSEN et al. (1982). Thus when the size of the control sample is large compared to the sample of cases, we may (approximately) compute the test statistic as if we had a sample of size n from the original Markov chain model. This becomes an exact computation when the functions $\bar{P}_h$ are known (i.e. we take $\lambda = 0$).

REMARK 3.5. $\hat{Z}\hat{C}^{-}\hat{Z}^T$ may be computed by deleting the last component of $\hat{Z}$ and the last row and column of $\hat{C}$, to give $\hat{Z}_0$ and $\hat{C}_0$ say, and then using the relation

$$\hat{Z}\hat{C}^{-}\hat{Z}^T = \hat{Z}_0 \hat{C}_0^{-1} \hat{Z}_0^T.$$

3.E Differential mortality

Let us no longer assume non-differential mortality as in Subsections 3.B,C, and D, but only that $\alpha_{h\mathcal{D}} \equiv \mu_0$ for all $h \in H$ and $\alpha_{i\mathcal{D}} \equiv \mu_1$ for all $i \in I$. For this situation we have by (3.3), the remark at the end of Section 2 and the obvious relation $P_{iI}(x,y) = \exp(-\int_x^y \mu_1(u)du)$ that

$$E N_{hi}^{(n)}(x) = \int_0^x \alpha_{hi}(s) \bar{P}_h(s) f(s) ds / P_I(\zeta),$$

where $f(s) = \exp\{-\int_0^s \mu_0(u)du - \int_s^\zeta \mu_1(u)du\}$ is independent of $h \in H$, $i \in I$. Hence

$$(3.20)\qquad C_{hi}(x) = \frac{\bar{P}_H(\zeta)}{P_I(\zeta)} \int_0^x [\alpha_{hi}(s)f(s)/\bar{P}_{hH}(s,\zeta)]ds = \int_0^x \frac{d\bar{E}N_{hi}(s)}{P_h^*(s)} \quad ,$$

where $P_h^*(x) = P\{S(x) = h \,|\, S(\zeta) \in H\} = \bar{P}_h(x)\bar{P}_{hH}(x,\zeta)/\bar{P}_H(\zeta)$ for $x \le \zeta$, $h \in H$. In this subsection we will assume that the disease is a rare one, so that $\bar{P}_I(\zeta) \approx 1$ and $\bar{P}_{hH}(x,\zeta) \approx 1$ and therefore

$$C_{hi}(x) \approx \int_0^x \alpha_{hi}(s)f(s)ds/P_I(\zeta).$$

Even if the C_{hi}'s are not proportional to the integrated intensities it will be useful to get estimates for these quantities, e.g. for assessing the *relative* size of the intensities and for a graphical check of the assumption of proportional intensities.

By (3.20) it is seen that one now should draw a random sample of m controls among those alive and non-diseased at age ζ. We assume that this sample of controls is "representative" in the same manner as indicated in Subsection 3.B, and let $Y_h^{(m)}(x)$ and $Y_{hk}^{(m)}(x,y)$ refer to these control individuals. Then estimators for the C_{hi}'s are

$$(3.21)\qquad \hat{C}_{hi}(x) = \int_0^x \frac{dN_{hi}^{(n)}(s)}{Y_h^{(m)}(s)}$$

for $h \in H$, $i \in I$. Their properties follow by substituting C_{hi} for B_{hi} and $P_h^*(x)$ and $P_{hk}^*(x,y)$ for $\bar{P}_h(x)$ and $\bar{P}_{hk}(x,y)$ in Theorem 3.1. Moreover, to test the hypothesis that for a given set $R \subset H \times I$, $C_{hi} = C$ for all $(h,i) \in R$ (which is (almost) the same as the equality of A_{hi} for $(h,i) \in R$), we may use the results of Theorem 3.2 with the obvious modifications.

4. SOME OTHER SAMPLING FRAMES

In this section we will have a brief look at some other sampling frames for the cases and the controls. We will first consider sampling from a specific cohort. Thereafter we will comment upon some problems related to the sampling of new or incident cases of the disease in question.

Let us assume that the n cases are a sample of the individuals in a given cohort who are dead with the disease at a given "age" ζ; sampled in such a way that all of these individuals have the same probability of being

selected. Moreover, assume that information on their individual exposure and disease histories may be obtained, e.g. from a national health register. To model this situation we split the set $\mathcal{D}$ of "dead" states into a disjoint union $\mathcal{D}_H \cup \mathcal{D}_I$, where $\mathcal{D}_H$ contains the "dead" states for people who never get the disease, and $\mathcal{D}_I$ contains the "dead" states for diseased individuals. One example of such a model is the extension of the model of Fig.1 shown in Fig.2. Here $\mathcal{D}_H = \{\dagger_H\}$ and $\mathcal{D}_I = \{\dagger_I\}$. Now, our sampling scheme is such that

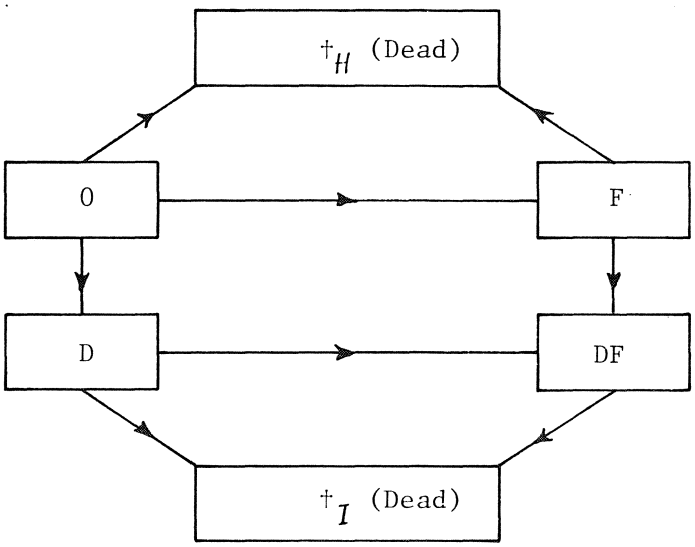

Fig.2 The Markov model of Fig.1 extended with two death states

all cases will be in one of the states in $\mathcal{D}_I$ at "age" ζ. Hence, the observations (from the cases) come from the Markov chain obtained by conditioning on being in $\mathcal{D}_I$ at "age" ζ.

Let $N_{hi}^{(n)}(x)$ denote 1/n times the number of transitions directly from state $h \in H$ to state $i \in I$ among the n cases in the age span $[0,x]$. Then, by calculations similar to those giving (3.3),
$EN_{hi}^{(n)}(x) = \int_0^x \alpha_{hi}(s)P_h(s)P_{i\mathcal{D}_I}(s,\zeta)ds/P_{\mathcal{D}_I}(\zeta)$. If we assume that $\alpha_{h\mathcal{D}_H} \equiv \mu_0$ for all $h \in H$ and $\alpha_{i\mathcal{D}_I} \equiv \mu_1$ for all $i \in I$, then as in Subsection 3.E

$$EN_{hi}^{(n)}(x) = \int_0^x \alpha_{hi}(s)\bar{P}_h(s)g(s)ds/P_{\mathcal{D}_I}(\zeta),$$

where

$$g(s) = \exp(-\int_0^s \mu_0(u)du)(1-\exp(-\int_s^\zeta \mu_1(u)du)),$$

so that

$$D_{hi}(x) = \frac{\bar{P}_H(\zeta)}{P_{\mathcal{D}_I}(\zeta)} \int_0^x [\alpha_{hi}(s)g(s)/\bar{P}_{hH}(s,\zeta)]ds = \int_0^x \frac{dEN_{hi}(s)}{P_h^*(s)} .$$

Here P_h^* is defined just below (3.20). Hence, D_{hi} has the same structure as C_{hi} in Subsection 3.E and may be estimated and interpreted more or less as described there.

Note that if the cases consist of a sample of those alive or dead with the disease at age ζ the analysis above goes through with $\mathcal{D}_I$ replaced by $I \cup \mathcal{D}_I$ and $g(s) = \exp(-\int_0^s \mu_0(u)du)$. Moreover no assumptions on the $\alpha_{i\mathcal{D}_I}$'s for $i \in I$ are necessary.

A quite common procedure in practice is to collect a sample of *new cases* of the disease in question together with a sample of controls selected randomly among those alive without the disease (irrespectively of age). In such a situation the cases as well as the controls come from various birth cohorts, possibly of varying sizes. We are presently investigating the possibilities of applying our techniques to this situation as well, and there seem to be no insurmountable difficulties in this. We intend to report on these results later.

5. AN ILLUSTRATION: PUSTULOSIS PALMO-PLANTARIS AND MENOPAUSE

To illustrate the methods of this paper we will reanalyse in part the data in Section 3 of AALEN et al. (1980). This example concerns the possible influence of menopausal hormonal changes or similar artificially induced changes in ovarian function ("induced menopause") on the intensity of the outbreak of the chronical skin disease *pustulosis palmo-plantaris*; i.e. the medical question is whether menopause is a "risk factor" for this skin disease. The present example is in fact too simple to really motivate the rather complicated procedures we propose. Nevertheless it should illustrate

the type of results one may obtain by our methods.

Since pustulosis palmo-plantaris is a relatively rare disease, with a reported prevalence of 0.05 per cent in a Swedish study (HELLGREN & MOBACKEN, 1971), the data of AALEN et al. (1980) are based on interviews with 85 females at the Department of Dermatovenerology, Finsen Institute, Copenhagen, *all suffering from the disease in question*. No controls were sampled, and their analysis was performed separately for natural and induced menopause. However, from "The cardiovascular disease study in Norwegian counties", performed by the National Mass Radiography Service, Oslo, they obtained information on the age distribution of menopause for ages between 35 and 50 years. AALEN et al. (1980) used this information on onset of menopause only to determine the intensity with which menopause occurs. Here we will show how the Norwegian menopause data also may be used as "controls". To this end we omit the Copenhagen induced menopause women to make the Norwegian and Copenhagen female populations more comparable. However, it should still be kept in mind that the two female populations we consider may have different age distributions for natural menopause, which may make it somewhat misleading to use the Norwegian data as controls. Nevertheless the main patterns should be the same, so that the data are appropriate for our mainly illustrative purposes.

The Norwegian menopause data are given in Table 1 together with the number of occurrences of pustulosis palmo-plantaris before and after natural menopause in the Copenhagen female population (extracted from Table 1 of AALEN et al., 1980). It is assumed that for the three women who reported menopause and the outbreak of the disease to happen simultaneously, the disease was the last event. Since the Norwegian-menopause data are based on interviews with around 25000 women we assume below that the age-specific prevalence of menopause is known and equal to the numbers in Table 1.

We will analyse the data by means of the Markov chain model of Fig. 1, where F now stands for menopause. As explained by AALEN et al. (1980) none of the sampling frames discussed in Sections 3 and 4 of this paper give a satisfactory description of the process of data selection from the diseased population. A more appropriate description is to assume that any given patient has a fixed intensity of being sampled per unit of time in which the patient is diseased and still alive. The combined biological and sampling

Table 1. Age distribution of menopause for the Norwegian female population and occurrences of pustulosis palmo-plantaris in the Copenhagen sample before and after menopause [a)]

Age	Prevalence of menopause	Occurrences of pustuloris palmo-plantaris	
		Before menopause	After menopause
35	.024	1	-
36	.029	2	-
37	.036	1	-
38	.021	1	-
39	.055	-	-
40	.056	5	-
41	.076	1	1
42	.099	-	-
43	.112	1	1
44	.167	2	-
45	.224	2	2
46	.282	-	-
47	.364	-	1
48	.449	1	2
49	.579	-	3

a) The data include only natural menopause. For the three women who reported the same age for menopause and the disease, the disease is assumed to be the last event.

process may then be approximated by the model described in Section 2C of AALEN et al. (1980). If we assume *non-differential* mortality it then follows that we may think of the data on the diseased individuals as coming from a Markov chain model with intensities $\alpha_{hi}(x)f(x)/\pi_h(x)$ for $h \in \{0,F\}$, $i \in \{D,DF\}$, where $f(x)$ only depends on x, the sampling intensity λ, and the common force of mortality μ, while $\pi_h(x)$ denotes the probability that a female in state h at age x will eventually get sampled (after being diseased, cf. AALEN et al., 1980, Section 2C).

By an argument similar to the ones given in Section 3 of this paper it

then follows that

$$D_{hi}(x) = \int_0^x \alpha_{hi}(s)f(s)p(s)ds/\pi_0(0),$$

with $p(x) = \exp(-\int_0^x \mu(s)ds)$, is equal to $\int_0^x (d\bar{E}N_{hi}^{(n)}(s)/\bar{P}_h(s))$, where $N_{hi}^{(n)}(x)$ is $1/n$ times the number of transitions directly from state $h \in \{0,F\}$ to state $i \in \{D,DF\}$ reported by the $n = 66$ female patients in the age span from 0 to x years. (We have omitted the 19 induced menopause women.)
Hence $D_{hi}(35,x) = D_{hi}(x) - D_{hi}(35)$ may be estimated by

$$\hat{D}_{hi}(35,x) = \int_{35}^x \frac{dN_{hi}^{(n)}(s)}{\bar{P}_h(s)},$$

where $\bar{P}_F(\cdot) = 1 - \bar{P}_0(\cdot)$ is given in the second column of Table 1 and $ndN_{0D}^{(n)}(\cdot)$ and $ndN_{F,DF}^{(n)}(\cdot)$ in the third and fourth column, respectively. The statistical properties of these estimators follow from Theorem 3.1 and Remark 3.1 after the appropriate substitutions are made.

Plots of $\hat{D}_{0D}(35,x)$ and $\hat{D}_{F,DF}(35,x)$ are given in Fig.3. The plots give the impression that $\alpha_{F,DF}$ is greater that α_{0D} at least for ages above 40 years. To test the hypothesis H_0: $\alpha_{0D}(x) = \alpha_{F,DF}(x)$ for $35 \le x \le 50$, or equivalently $D_{0D}(35,x) = D_{F,DF}(35,x)$ for $35 \le x \le 50$, we may use the test statistic (cf. Remark 3.4)

$$S = \hat{Z}_{F,DF}(49) - \hat{Z}_{F,DF}(34) =$$

$$= n^{\frac{1}{2}}(N_{F,DF}^{(n)}(49) - N_{F,DF}^{(n)}(34) - \int_{35}^{49} \bar{P}_F(s)dN_{\cdot}^{(n)}(s)),$$

where $N_{\cdot}^{(n)} = N_{0D}^{(n)} + N_{F,DF}^{(n)}$. (Thus we take the time interval $(34,49]$ instead of $[0,\zeta]$ in (3.18).) An estimate of its variance, putting $m = \infty$ in (3.19), is

$$V = \hat{C}_{(F,DF),(F,DF)} = \int_{35}^{49} \bar{P}_F(s)\bar{P}_0(s)dN_{\cdot}^{(n)}(s).$$

By Theorem 3.2 it follows that $SV^{-\frac{1}{2}}$ is asymptotically normal $(0,1)$ under the hypothesis, as the number of patients increases to infinity. Now, $SV^{-\frac{1}{2}}$ takes

the significant value 2.45. (If for the three women who reported the same age for menopause and the disease, the disease is assumed to be the first event, we instead get the value 0.81 corresponding to a (one-sided) significance probability of 20.9%.) Thus, this new analysis seems to confirm the conclusions of AALEN et al. (1980) that menopause is a "risk factor" for pustulosis palmo-plantaris. In addition the new analysis gives information on the size of the increase in the intensity of the disease after the occurrence of menopause (comparing the slopes in Figure 3, it appears that the intensity is approximately doubled).

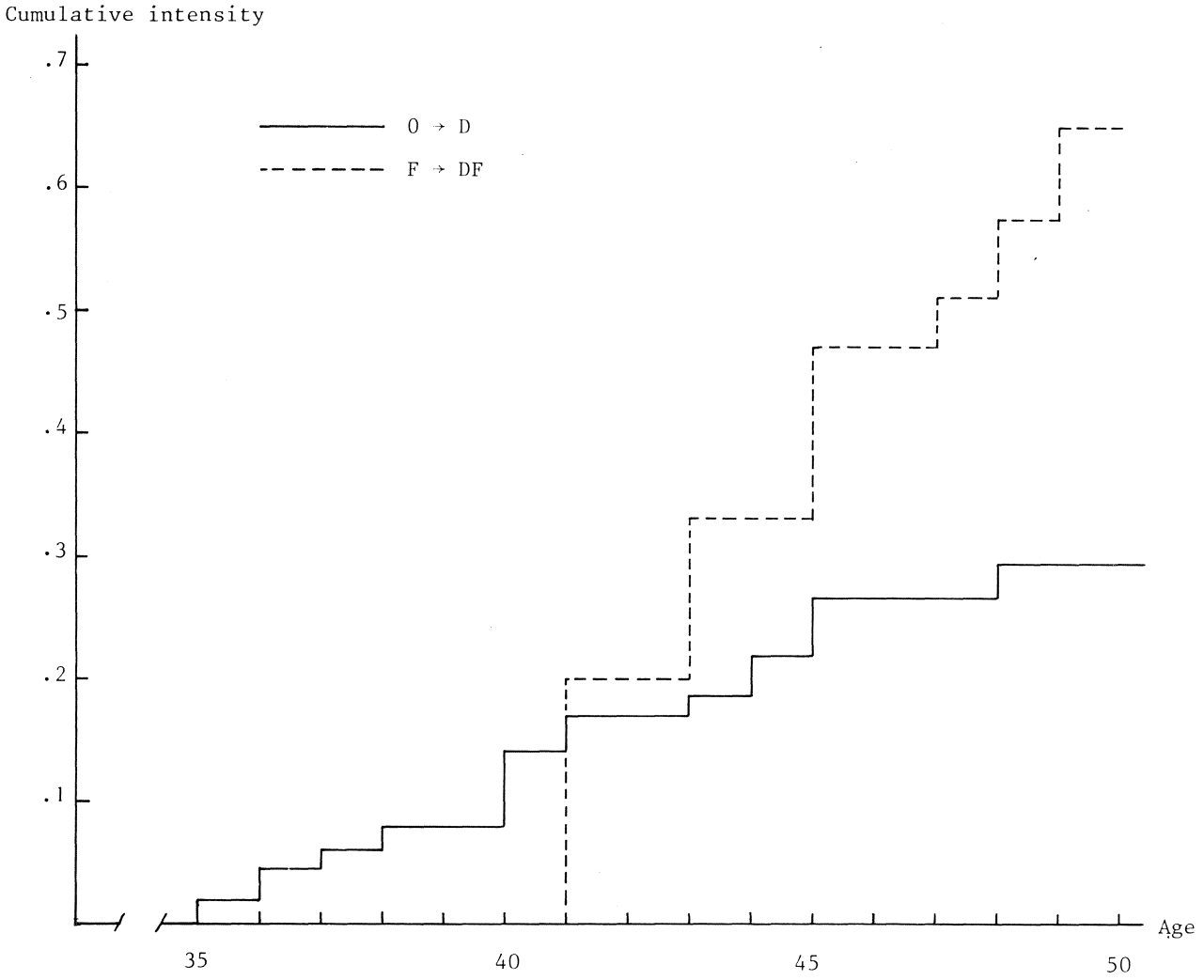

Fig. 3. Estimated cumulative intensities for occurrence of pustulosis palmo-plantaris (D) before menopause (O) and after menopause (F).

APPENDIX 1

TECHNICAL LEMMAS FOR PROOF OF THEOREMS 3.1 AND 3.2

In order to facilitate the extension to discontinuous cumulative intensities in Appendix 2, we are careful in this appendix to distinguish left continuous and right continuous versions of processes and functions, and we draw attention to the single point where the extension to the discontinuous case is non trivial.

1.A. Weak convergence of $\{m^{\frac{1}{2}}(Y_h^{(m)} - EY_h) : h \in L\}$

Recall that $Y_h^{(m)}(x)$ is the fraction of the m control individuals who are in state h just before time $x \in [0,\zeta]$, where each individual's path is a Markov chain on $L = H \cup I$ with cumulative intensities A_{ij}, $i \neq j \in L$. The m individuals travel independently of one another on this chain, and each has the same initial distribution over L. Clearly the finite dimensional distributions of the multivariate process $\{m^{\frac{1}{2}}(Y_h^{(m)} - EY_h) : h \in L\}$ converge to multivariate normal distributions as $m \to \infty$. In order to show that this process converges in distribution in $(D[0,\zeta])^r$, where $r = \#L$ and elements of $D[0,\zeta]$ are here defined to be *left* continuous with *right* hand limits, it suffices to show that left continuous functions F_h and G_h exist such that, for $0 \leq s \leq t \leq u \leq \zeta$ and $h \in L$

(i) $E((Y_h(u)-Y_h(t))^2) \leq F_h(u) - F_h(t)$

(ii) $E((Y_h(u)-Y_h(t))^2(Y_h(t)-Y_h(s))^2) \leq (G_h(u)-G_h(t))(G_h(t)-G_h(s))$

where $Y_h = Y_h^{(1)}$ (cf. BILLINGSLEY (1968) Theorem 15.6 and the remarks on page 133, and HAHN (1978) Theorem 2). If F_h and G_h are actually continuous, then the limiting process has paths which are a.s. continuous too.

From now on we restrict attention to the case m = 1. For $i,j \in L$, $i \neq j$, let $N_{ij}(x)$ be the number of transitions directly from i to j in the time interval $[0,x]$. Write $N_{i\cdot}$, $N_{\cdot i}$, $A_{i\cdot}$, $A_{\cdot i}$ for $\sum_{j \neq i} N_{ij}$ etc. Note that for $t < u$

$$(Y_h(u)-Y_h(t))^2 = |Y_h(u)-Y_h(t)| \leq N_{h\cdot}(u-)-N_{h\cdot}(t-) + N_{\cdot h}(u-)-N_{\cdot h}(t-)A_{i\cdot}$$

since $Y_h(u)$ and $Y_h(t)$ can only differ (and then only by the amount ± 1)

if there has been a transition out of or into h in the interval [t,u). By AALEN's (1978) theory of counting processes (extended to possibly discontinuous A_{ij} by GILL, 1980b), conditional on the state just before time t, we have

$$N_{ij}(\cdot) - N_{ij}(t-) - \int_{[t,\cdot]} Y_i(u)\,dA_{ij}(u)$$

is a martingale on $[t,\zeta]$ for all $i,j \in L$. Thus in particular we have

$$(A.1.2)\qquad E[N_{ij}(u-)-N_{ij}(t-) \mid Y_k(t):k \in L] = E\Big[\int_{[t,u)} Y_i(u)\,dA_{ij}(u)\Big]$$

$$\leq A_{ij}(u-) - A_{ij}(t-).$$

Thus we obtain from (A.1.1) and (A.1.2)

$$E((Y_h(u)-Y_h(t))^2) \leq (A_{h\cdot}(u-)-A_{h\cdot}(t-))+(A_{\cdot h}(u-)-A_{\cdot h}(t-))$$

and

$$\begin{aligned} &E((Y_h(u)-Y_h(t))^2(Y_h(t)-Y_h(s))^2) \\ &\quad = E(E((Y_h(u)-Y_h(t))^2|Y_k(t):k \in L)E((Y_h(t)-Y_h(s))^2|Y_k(t):k \in L) \\ &\quad \leq (A_{h\cdot}(u-)-A_{h\cdot}(t-)+A_{\cdot h}(u-)-A_{\cdot h}(t-))\cdot E((Y_h(t)-Y_h(s))^2) \\ &\quad \leq (A_{h\cdot}(u-)-A_{h\cdot}(t-)+A_{\cdot h}(u-)-A_{\cdot h}(t-)) \\ &\qquad\qquad (A_{h\cdot}(t-)-A_{h\cdot}(s-)+A_{\cdot h}(t-)-A_{\cdot h}(s-)). \end{aligned}$$

Thus (i) and (ii) hold with $F_h(t) = G_h(t) = A_{\cdot h}(t-) + A_{h\cdot}(t-)$ for all t and h. Also we see that the limiting multivariate Gaussian process has continuous paths if each A_{ij} is continuous.

Note that from (A.1.1) we see that

$$\int_0^\zeta |dY_h^{(m)}(t)| \leq N_{h\cdot}^{(m)}(\zeta)+N_{\cdot h}^{(m)}(\zeta) \xrightarrow{P} EN_{h\cdot}(\zeta)+EN_{\cdot h}(\zeta)$$

$$\leq A_{h\cdot}(\zeta)+A_{\cdot h}(\zeta).$$

Therefore there exists a constant $C < \infty$ such that

$$I\left\{ \int_0^{\zeta} |dY_h^{(m)}(t)| \le C \right\} \xrightarrow{P} 1 \quad \text{as } m \to \infty.$$

1.B. Weak Convergence of Certain Functions of $\{Y_h^{(m)} : h \in H\}$ and $\{N_{hi}^{(n)} : (h,i) \in H \times I\}$

In order to concentrate on the important parts of the problem, we consider here a "one-dimensional" problem concerning $Y_h^{(m)}$ and $N_{hi}^{(n)}$ for a single $h \in H$ and $i \in I$, and we shall henceforth denote these processes by $Y^{(m)}$ and $N^{(n)}$. The only new difficulties in the multivariate case are notational ones.

We shall consider the problem of proving weak convergence of

$$\{n^{\frac{1}{2}}(\int_0^{\cdot} f(Y^{(m)}) dN^{(n)} - \int_0^{\cdot} f(EY) dEN)\}$$

where f is any sufficiently smooth function. The structure we need concerning the processes $Y^{(m)}$ and $N^{(n)}$, and the functions EY, EN and f, is contained in the following list of assumptions:

(i) As $n,m \to \infty$ in such a way that $n/m \to \lambda \in [0,\infty)$, $(n^{\frac{1}{2}}(N^{(n)}-EN), m^{\frac{1}{2}}(Y^{(m)}-EY)) \xrightarrow{\mathcal{D}} (U,V)$ in $(D[0,\zeta])^2$ (whose second component is considered to be the space of *left* continuous functions with *right* hand limits).

(ii) $\exists C < \infty$ s.t. $I\{\int_0^{\zeta} |dY^{(m)}| \le C\} \xrightarrow{P} 1$

(iii) $\int_0^{\zeta} |dN^{(n)}| < \infty$ almost surely for each n

(iv) $\int_0^{\zeta} |dEN| < \infty$, $\int_0^{\zeta} |dEY| < \infty$

(v) f has a continuous derivative f' on $[c_1-\varepsilon, c_2+\varepsilon]$ where ε, c_1 and c_2 satisfy $\varepsilon > 0$ and $c_1 \le EY(t) \le c_2$ for all $t \in [0,\zeta]$.

(vi) U and V have continuous sample paths with probability 1.

Under these assumptions we have

THEOREM A.1. *As* $n \to \infty$,

$$\left\{ n^{\frac{1}{2}} \left(\int_0^{\cdot} f(Y^{(m)}) dN^{(n)} - \int_0^{\cdot} f(EY) dEN \right) \right\} \xrightarrow{\mathcal{D}}$$

$$\left\{ \int_0^{\cdot} f(EY)dU + \lambda^{\frac{1}{2}} \int_0^{\cdot} f'(EY)VdEN \right\}$$

in $D[0,\zeta]$, *where the integral with respect to* U *is defined by formally integrating by parts. If it is also defined as a stochastic integral, then the two definitions coincide.*

PROOF. We closely follow BRESLOW & CROWLEY (1974). First we apply a Skorohod construction to the sequence of random elements of $(D[0,\zeta])^2 \times \mathbb{R}$:

$$\xi^{(n)} = \left\{ n^{\frac{1}{2}}(N^{(n)}-EN), m^{\frac{1}{2}}(Y^{(m)}-EY), I\left\{ \int_0^{\zeta} |dY^{(m)}| < C \right\} \right\}$$

which converges in distribution to $\xi = \{U,V,1\}$. Thus there exists a sample space with defined on it $\xi^{(n)'}$, $n = 1,2,\ldots$ and ξ' such that

$$\xi^{(n)'} \overset{\mathcal{D}}{=} \xi^{(n)} \quad \forall n$$

$$\xi' \overset{\mathcal{D}}{=} \xi$$

and

$$\xi^{(n)'} \xrightarrow{a.s.} \xi'$$

in the topology of $(D[0,\zeta])^2 \times \mathbb{R}$. Since U and V have continuous sample paths, we actually have convergence in the supremum norm (rather than just in the Skorohod metric) of the first two components of $\xi^{(n)'}$. (This is the one and only point at which continuity of the A_{ij}'s is crucial.) On this new sample space we shall prove almost sure convergence in the supremum norm of the functional of interest. This implies convergence in distribution (in $D[0,\zeta]$) in the original set-up.

So we now work with the stronger assumptions (i)' to (v)' obtained by replacing $\overset{\mathcal{D}}{\to}$ and $\overset{P}{\to}$ in (i) and (ii) by $\xrightarrow{a.s.}$ (w.r.t. to the supremum norm in (i)). Note first that by (v)', f' is bounded and uniformly continuous on $[c_1-\varepsilon, c_2+\varepsilon]$. Write $\|\cdot\|$ for the supremum norm on $[0,\zeta]$. On the event $\{Y^{(m)}(t) \in [c_1-\varepsilon, c_2+\varepsilon]\}$ we have by the mean value theorem

$$m^{\frac{1}{2}}(f(Y^{(m)}(t))-f(EY(t))) = f'(Y^{(m)*}(t))m^{\frac{1}{2}}(Y^{(m)}(t)-EY(t)),$$

where $Y^{(m)*}(t)$ lies between $Y^{(m)}(t)$ and $EY(t)$. Therefore by (i)', on the above mentioned event,

$$\|m^{\frac{1}{2}}(f(Y^{(m)})-f(EY))-f'(EY).V\|$$

$$\leq \|f'(Y^{(m)*})-f'(EY)\| \cdot \|m^{\frac{1}{2}}(Y^{(m)}-EY)\|$$

$$+ \|f'(EY)\| \cdot \|m^{\frac{1}{2}}(Y^{(m)}-EY)-V\| .$$

Since almost surely, for all large enough m we do have $Y^{(m)} \in [c_1-\varepsilon, c_2+\varepsilon]$, it follows that

$$\|m^{\frac{1}{2}}(f(Y^{(m)})-f(EY))-f'(EY)\cdot V\| \to 0$$

almost surely as $m \to \infty$.

Now we have

$$n^{\frac{1}{2}}\Big(\int_0^{\cdot} f(Y^{(m)})\,dN^{(n)} - \int_0^{\cdot} f(EY)\,dEN\Big)$$

$$= \int_0^{\cdot} f(Y^{(m)})\,d(n^{\frac{1}{2}}(N^{(n)}-EN)) + \Big(\frac{n}{m}\Big)^{\frac{1}{2}} \int_0^{\cdot} m^{\frac{1}{2}}(f(Y^{(m)})-f(EY))\,dEN$$

and the limiting process is a sum of two corresponding components. We see directly that

$$\Big\|\Big(\frac{n}{m}\Big)^{\frac{1}{2}} \int_0^{\cdot} m^{\frac{1}{2}}(f(Y^{(m)})-f(EY))\,dEN - \lambda^{\frac{1}{2}} \int_0^{\cdot} f'(EY)V\,dEN\Big\| \to 0 \text{ a.s.}$$

Moreover, we have on the event $\{Y^{(m)} \in [c_1-\varepsilon, c_2+\varepsilon]\}$ that $\int_0^{\zeta} |df(Y^{(m)})| \leq \|f'\| \int_0^{\zeta} |dY^{(m)}|$ and $\int_0^{\zeta} |df(EY)| \leq \|f'\| \int_0^{\zeta} |dEY|$ where $\|f'\| = \sup\{f'(y): y \in [c_1-\varepsilon, c_2+\varepsilon]\}$. Therefore by GILL (1980a) Lemma 5 (an abstract version of part of BRESLOW and CROWLEY, 1974, Theorem 4)

$$\Big\|\int_0^{\cdot} f(Y^{(m)})\,d(n^{\frac{1}{2}}(N^{(n)}-EN)) - \int_0^{\cdot} f(EY)\,dU\Big\| \to 0 \text{ a.s.},$$

which proves the stated weak convergence. The remaining part of the theorem

is a direct consequence of Lemma 5 in GILL (1980a). □

1.C. Consistency of estimators of asymptotic covariance structure

An analysis of the problem of establishing uniform consistency of the estimators of covariance functions in Theorem 3.1 and 3.2 show that the following two Lemmas on random signed measures in $\mathbb{R}^p$ cover all the difficult points. To motivate the Lemmas, we point out that we will apply Lemma 2 in Theorem 3.1, considering $m = m(n)$, by taking $p = 2$, $\mu^{(n)}$ to be the random signed measure defined by $\mu^{(n)}(dx_1,dx_2) = Y_{hk}^{(m)}(dx_1,dx_2)$, and $\nu^{(n)}$ to be the random measure defined by

$$\nu^{(n)}(dx_1,dx_2) = \frac{N_{hi}^{(n)}(dx_1)}{Y_h^{(m)}(x_1)^2} \frac{N_{k\ell}^{(n)}(dx_2)}{Y_k^{(m)}(x_2)^2} .$$

First we introduce some notation. For a measure μ on the Borel sets of $\mathbb{R}^p$ and points $x = (x_1,\ldots,x_p)$ and $y = (y_1,\ldots,y_p) \in \bar{\mathbb{R}}^p$ we define

$$(x,y] = \prod_{i=1}^{p} (x_i,y_i] \quad \text{(empty if } x_i > y_i \text{ for some } i)$$

$$(x,y) = \prod_{i=1}^{p} (x_i,y_i) \quad \text{(empty if } x_i \geq y_i \text{ for some } i)$$

$$\mu(x) = \mu((-\infty,x])$$

$$\mu_-(x) = \mu((-\infty,x))$$

(note that $\mu(x) \neq \mu(\{x\})$!). Define two norms on the space of measures by

$$\|\mu\|_\infty = \sup_{x\in\mathbb{R}_p} |\mu(x)| = \sup_{x\in\mathbb{R}_p} |\mu_-(x)|$$

and

$$\|\mu\|_V = \int_{x\in\mathbb{R}_p} |\mu(dx)| .$$

LEMMA 1. *Let $\mu_1,\mu_2,\ldots$ be an i.i.d. sequence of random measures on $\mathbb{R}^p$ and define $\mu^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.*

Suppose that $E\|\mu_i\|_V < \infty$ and define $\mu = E\mu_i$. Then $\|\mu^{(n)} - \mu\|_\infty \to 0$ a.s. as $n \to \infty$ and

$$\limsup_{n\to\infty} \int_{x\in\mathbb{R}^p} |\mu^{(n)}(dx)| < \infty \quad \text{a.s.}$$

PROOF. This result can be established by generalizing many of the standard proofs of the Glivenko-Cantelli lemma. In particular, the proof of RAO's (1962) Theorem 7.2 lends itself very easily to this extension. □

LEMMA 2. *Let* $(\mu^{(n)}), (\nu^{(n)})$ *be two sequences of random measures and* μ, ν *be two fixed measures on* $\mathbb{R}^p$ *such that*

$$\|\mu^{(n)} - \mu\|_\infty \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty$$

$$\|\nu^{(n)} - \nu\|_\infty \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty$$

$$\lim_{C\uparrow\infty} \limsup_{n\to\infty} P[\|\nu^{(n)}\|_V > C] = 0$$

$$\|\mu\|_V < \infty, \|\nu\|_V < \infty.$$

Define $\xi^{(n)}$ and ξ by $\xi^{(n)}(dx) = \mu_-^{(n)}(x)\nu^{(n)}(dx)$, $\xi(dx) = \mu_-(x)\nu(dx)$. Then

$$\|\xi^{(n)} - \xi\|_\infty \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty$$

and

$$\lim_{C\uparrow\infty} \limsup_{n\to\infty} P[\|\xi^{(n)}\|_V > C] = 0.$$

PROOF. The second assertion is easy to verify so we only consider the first one. Note the relation

$$\int_{y\in(-\infty,x]} \mu_-(y)\nu(dy) = \int_{y\in(-\infty,x]} \left(\int_{z\in(-\infty,y)} \mu(dz) \right) \nu(dy)$$

$$= \int_{z\in(-\infty,x)} \left(\int_{y\in(z,x]} \nu(dy) \right) \mu(dz)$$

$$= \int_{z\in(-\infty,x)} \sum_{\omega\in P(z,x)} (-1)^{s(\omega,z)} \nu(\omega)\, \mu(dz),$$

where $P(z,x) = \{\omega \in \mathbb{R}^p : \omega_i = z_i \text{ or } x_i \text{ for each } i\}$ is the set of 2^p corners of the hyper rectangle $(z,x]$, and $s(\omega,x) = p$ - (number of components of ω equal to the corresponding component of x). Thus

$$\|\xi\|_\infty \le 2^p \|\mu\|_V \|\nu\|_\infty .$$

Since

$$\xi^{(n)}(x) - \zeta(x) = \int_{y\in(-\infty,x]} (\mu_-^{(n)} - \mu)(y)\, \nu^{(n)}(dy)$$

$$+ \int_{y\in(-\infty,x]} \mu_-(y)\, (\nu^{(n)} - \nu)(dy)$$

we obtain

$$\|\xi^{(n)} - \xi\|_\infty \le \|\mu^{(n)} - \mu\|_\infty \|\nu^{(n)}\|_V + 2^p \|\mu\|_V \|\nu^{(n)} - \nu\|_\infty$$

and the required result follows immediately. □

Applying Lemma 2 to Theorem 3.1 as indicated above, we must verify the conditions pertaining to μ and $\mu^{(n)}$ where $\mu^{(n)}$ is the measure generated by $Y_{hk}^{(m)}$. These conditions will follow from Lemma 1 since $Y_{hk}^{(m)}$ is $1/m$ times a sum of m i.i.d. random measures. Therefore we must check that

$$E \int_{x\in[0,\zeta)^2} |Y_{hk}^{(1)}(dx_1, dx_2)| < \infty$$

(we set $\mu^{(n)}, \nu^{(n)}$ zero outside $[0,\zeta]^2$).

Now $Y_{hk}^{(1)}(x_1,x_2) = Y_h^{(1)}(x_1) Y_k^{(1)}(x_2)$. Also we have

$$\int_{x_1 \in [0,\zeta)} |Y_h^{(1)}(dx_1)| = Y_h^{(1)}(0) + N_{h\cdot}^{(1)}(\zeta-) + N_{\cdot h}^{(1)}(\zeta-);$$

similarly for state k. Thus it is sufficient to show that

$$E[N_{ij}(\zeta)^2] < \infty \quad \text{for all} \quad i,j \in L.$$

Now by AALEN (1978) (extended to possibly discontinuous A_{ij} by GILL, 1980b)

$$(A.1.3) \qquad M_{ij}(\cdot) = N_{ij}(\cdot) - \int_0^{\cdot} Y_i(s)\,dA_{ij}(s)$$

and

$$(A.1.4) \qquad M_{ij}(\cdot)^2 - \int_0^{\cdot} Y_i(s)(1-\Delta A_{ij}(s))\,dA_{ij}(s)$$

are zero mean martingales on $[0,\zeta]$, where $\Delta A_{ij}(s) = A_{ij}(s) - A_{ij}(s-) \leq 1$ (see Appendix 2). Thus we have from (A.1.3)

$$E[N_{ij}(\zeta)] \leq A_{ij}(\zeta)$$

and from (A.1.4)

$$E[N_{ij}(\zeta)^2 - 2N_{ij}(\zeta)\int_0^{\zeta} Y_i(s)\,dA_{ij}(s) + (\int_0^{\zeta} Y_i(s)A_{ij}(s))^2)$$

$$= E[\int_0^{\zeta} Y_i(s)(1-\Delta A_{ij}(s))\,dA_{ij}(s)] \leq A_{ij}(\zeta).$$

This gives

$$E[N_{ij}(\zeta)^2] \leq 2A_{ij}(\zeta)E[N_{ij}(\zeta)] + A_{ij}(\zeta) \leq A_{ij}(\zeta)(2A_{ij}(\zeta)+1) < \infty.$$

The rest of the application to Theorems 3.1 and 3.2 is left to the reader. (Note that the conditions on the derivative of L in Theorem 3.2 will be needed when considering the covariance estimator (3.17).)

APPENDIX 2

GENERALIZATION TO CUMULATIVE INTENSITIES

We mentioned previously that all the results of this paper hold under the simple assumption that cumulative intensities A_{ij} exist; they need not be continuous, let alone absolutely continuous. In this appendix we explain what we mean precisely by cumulative intensities when the α_{ij}'s do not exist, and we sketch how the extension may be carried out. Finally we discuss the importance of such an extension.

Consider then a non-homogeneous, time continuous Markov chain $\{S(x):x \geq 0\}$ with finite state space J. We suppose that the sample paths of S are piecewise constant and right continuous, and that there are only a finite number of jumps in any finite time interval. Then S is equivalently described by random jump times $0 = T_0 < T_1 < T_2 < \ldots$ such that $T_i \uparrow \infty$ ($T_i = T_{i+1}$ if $T_i = \infty$) and random states $I_0, I_1, \ldots$ such that $S(x) = I_j$ throughout the interval $[T_j, T_{j+1})$, $j = 0,1,2,\ldots, T_j < \infty$.

Cumulative intensities are functions $A_{ij}: [0,\infty) \to [0,\infty)$, $i \neq j$, which are nondecreasing, right continuous, and zero at time zero ($A_{ij}(0) = 0$). Defining $A_{i\cdot} = \sum_{j \neq i} A_{ij}$, we also suppose that

$$\Delta A_{i\cdot}(x) = A_{i\cdot}(x) - A_{i\cdot}(x-) \leq 1 \text{ for all } x.$$

Then the functions A_{ij} are the cumulative intensities of the Markov chain S if

$$(1) \qquad P[T_{h+1} > v \mid T_h = u \text{ and } I_h = i] = \prod_{t \in (u,v]} (1 - dA_{i\cdot}(t))$$

and

$$(2) \qquad P[I_{h+1} = j \mid T_{h+1} = v] = \frac{dA_{ij}}{dA_{i\cdot}}(v)$$

for all $i,j \in J$, $i \neq j$, $0 \leq u < v$, and $h = 0,1,\ldots$.

By JACOBSEN (1972) the distribution of S is determined by the initial distribution (of S(0)) over J) and by the A_{ij}'s via (1) and (2): property (1)

gives the distribution of the sojourn time in a particular state, given the time of entry; property (2) gives the distribution over J of the new state given that a jump out of state j occurs at time v. Conversely, most Markov chains posses cumulative intensities. (Let T(u) be the time of the first transition in (u,∞). Then we must assume that if for any u and i the distribution of T(u) given S(u) = i has bounded support in (u,∞), then its support terminates in an atom of positive probability. In other words if $0 \leq u < v < \infty$ and $i \in J$ exist such that $P[T(u) \leq v \mid S(u) = i] = 1$, then there exists $v' > u$ such that

$$P[T(u) < v' \mid S(u) = i] < 1 \text{ and } P[T(u) \leq v' \mid S(u) = i] = 1.)$$

Two special cases are included in this set-up:
(i) When integrable forces of transition or intensities α_{ij} defined by

$$\alpha_{ij}(s) = \frac{\partial}{\partial t} P_{ij}(s,t)\Big|_{t=s}$$

exist, then $A_{ij}(t) = \int_0^t \alpha_{ij}(s)ds$, and the probabilities (1) and (2) are equal to $\exp(-\int_u^v \alpha_{i\cdot}(s)ds)$ and $\alpha_{ij}(v)/\alpha_{i\cdot}(v)$ respectively, where $\alpha_{i\cdot} = \sum_{j\neq i} \alpha_{ij}$.
(ii) When the Markov process S is actually a discrete time Markov chain, i.e. the jump times T_k are integer valued, then A_{ij} is constant on each interval $[t-1,t)$, t an integer, and

$$P_{ij}(t-1,t) = P_{ij}(t-,t) = \Delta A_{ij}(t),$$

for integer t. Probabilities (1) and (2) now become $\prod_{t\in(u,v]} (1-\Delta A_{i\cdot}(t))$ and $\Delta A_{ij}(v)/\Delta A_{i\cdot}(v)$ respectively.

We now discuss the extension of our results from absolutely continuous A_{ij}'s (example (i)) to arbitrary A_{ij}'s. We have appealed to results of HOEM (1969) and AALEN (1978) which in these papers are stated for the absolutely continuous case (in fact, continuity assumptions on the α_{ij}'s are also made). However, by the results of JACOBSEN (1972) and GILL (1980b) respectively, all these results can be immediately transferred to the case of arbitrary A_{ij}'s.

Given this fact, all our derivations go through in the case of continuous A_{ij}'s just as in the case of absolutely continuous A_{ij}'s.

When we extend to A_{ij}'s with jumps, only two problems arise, one of which is purely notational. When transitions can occur at fixed times with positive probability, the function $\bar{P}_i(x) = P\ [S(x) = i \mid S(\zeta) \notin \mathcal{D}]$ of Section 3 is no longer continuous in x and y, but only right continuous. However we defined $Y_h^{(m)}(x)$ as the fraction of control individuals in state h just before time x; thus

$$E[Y_h^{(m)}(x)] = \bar{P}_h(x-).$$

This means that in Theorems 3.1 and 3.2 we must redefine $\bar{P}_h$ and $\bar{P}_{hk}$ by

$$\bar{P}_h(x) = P[S(x-) = h \mid S(\zeta) \notin \mathcal{D}]$$

$$\bar{P}_{hk}(x,y) = P[S(y-) = k \mid S(x-) = h,\ S(\zeta) \notin \mathcal{D}].$$

The second problem unfortunately is highly technical. In Appendix 1.B we needed condition (vi), on continuity of the sample paths of U and V, in order to carry out a Skorohod construction with respect to the supremum norm rather than just with respect to the Skorohod metric for $D[0,\zeta]$. This difficulty can be overcome in the discontinuous case by inserting small intervals at each jump point of the A_{ij}'s, linearly interpolating the $Y_i^{(m)}$ and $N_{ij}^{(n)}$ processes across each interval, and proving weak convergence with a continuous limiting processes on the extended time axis. This technique is shown to work for a similar problem in GILL (1980a), and we do not wish to repeat the details here.

Finally we discuss the importance of such extensions. In the first place there is a clear mathematical importance, since the assertions of e.g. Theorem 3.1 can be made without assuming any smoothness of the cumulative intensities. It would be very unsatisfactory if smoothness assumptions had to be made. There is some practical importance too. In the literature one often comes accross the claim that for practical purposes it is suffcient to work with continuous intensities. In practice however one cannot *empirically* discern a smooth A_{ij} for a nonsmooth A_{ij}. Thus we

rather have exactly the opposite situation: because the theorems hold in the nonsmooth as well as in the smooth case, it does not matter that we imagine intensities as being continuous functions.

Acknowledgements

We want to thank the National Mass Radiography Service, Oslo, for providing the Norwegian menopause data. Ørnulf Borgan was supported by the Norwegian Research Council for Science and the Humanities and by the Association of Norwegian Insurance Companies.

REFERENCES

AALEN, O.O. (1978), *Nonparametric inference for a family of counting processes*, Ann. Statist. 6, 701-726.

AALEN, O.O., Ø. BORGAN, N. KEIDING and J. THORMANN (1980), *Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring*, Scand. J. Statist. 7, 161-171.

ANDERSEN, P.K., Ø. BORGAN, R.D. GILL and N. KEIDING (1982), *Linear nonparametric tests for comparison of counting processes, with applications to censored survival data*, To appear in Int. Stat. Rev. 50.

BILLINGSLEY, P. (1968), *Convergence of probability measures*, Wiley, New York.

BOOS, D.D. and R.J. SERFLING, (1980), *A note on differentials and the CLT and LIL for statistical functions, with application to M-estimators*, Ann. Statist. 8, 618-624.

BORGAN, Ø. (1980), *Applications of non-homogeneous Markov chains to medical studies. Nonparametric analysis for prospective and retrospective data*, Res. Report 8/80, Inst. of Math., Univ. of Oslo. Has appeared as pp.102-115 in Explorative Datenanalyse, Frühjahrstagung, München, 1980, Proceedings, eds. N. Victor, W. Lehmacher and W. van Eimeren, Springer's series "Medizinische Informatik und Statistik", vol. 26.

BRESLOW, N.E. and J. CROWLEY, (1974), *A large sample study of the life table and product limit estimates under random censorship*, Ann. Statist. 2, 437-453.

BRESLOW, N.E. and N.E. DAY, (1980), *Statistical methods in cancer research*, vol. 1: *The analysis of case-control studies*, International agency for research on cancer, Lyon.

COHEN, J.E., (1972), *When does a leaky compartment model appear to have no leaks?* Theor. Pop. Biol. 3, 404-405.

CORNFIELD, J., (1951), *A method of estimating comparative rates from clinical data, applications to cancer of the lung, breast and cervix,* J. Nat. Cancer Inst. 11, 1269-1275.

CORNFIELD, J., (1956), *A statistical problem arising from retrospective studies,* Proceedings of the third Berkeley Symposium on mathematical statistics and probability, Univ. of California press 4, 135-148.

CORNFIELD, J. and W. HAENZEL, (1960), *Some aspects of retrospective studies,* J. Chron. Dis. 11, 523-534.

DOOB, J.L. (1953), *Stochastic processes*, Wiley, New York.

GILL, R.D., (1980a), *Nonparametric estimation based on censored observations of a Markov renewal process,* Z. Wahrscheinlichkeitstheorie verw. Gebiete 53, 97-116.

GILL, R.D., (1980b), *Censoring and stochastic integrals,* MC Tract 124, Mathematical Centre, Amsterdam.

HAHN, M., (1978), *Central limit theorems in* D[0,1]. Z. Wahrscheinlichkeitstheorie verw. Gebiete 44, 89-101.

HOEM, J.M., (1969), *Purged and partial Markov chains,* Skand. Akt. Tidskr. 52, 147-155.

HOEM, J.M., (1972), *Inhomogeneous semi-Markov processes, select actuarial tables, and duration dependence in demography*, in Population Dynamics (ed. T.N.E. Greville), pp. 251-296, Academic Press, New York.

JACOBSEN, M., (1972), *A characterization of minimal Markov jump processes,* Z. Wahrscheinlichkeitstheorie verw. Gebiete 23, 32-46.

LILIENFELD, A.M. and D.E. LILIENFELD, (1979), *A century of case-control studies: progress?* J. Chron. Dis. 32, 5-13.

MANTEL, N. and W. HAENZEL, (1959), *Statistical aspects of the analysis of data from retrospective studies of disease*, J. Nat. Cancer Inst. 22, 719-748.

MEYER, P.A., (1976), *Un cours sur les integrales stochastique*, pp.245-400 in Séminaire de Probabilités X, Lecture Notes in Mathematics 511, Springer Verlag, Berlin.

MIETTINEN, O., (1976), *Estimability and estimation in case-referent studies*, Amer. J. Epidem. 103, 226-235.

RAO, R.R., (1962), *Relations between weak and uniform convergence of measures with applications*, Ann. Math. Statist. 33, 659-680.

SERFLING, R.J., (1980), *Approximation theorems of mathematical statistics*, Wiley, New York.